

Summer 8-2020

Bootstrapping Web Archive Collections From Micro-Collections in Social Media

Alexander C. Nwala
Old Dominion University, alexandernwala@gmail.com

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_etds



Part of the [Databases and Information Systems Commons](#), [Library and Information Science Commons](#), and the [Social Media Commons](#)

Recommended Citation

Nwala, Alexander C.. "Bootstrapping Web Archive Collections From Micro-Collections in Social Media" (2020). Doctor of Philosophy (PhD), Dissertation, Computer Science, Old Dominion University, DOI: 10.25777/ez78-cb43
https://digitalcommons.odu.edu/computerscience_etds/124

This Dissertation is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

**BOOTSTRAPPING WEB ARCHIVE COLLECTIONS FROM
MICRO-COLLECTIONS IN SOCIAL MEDIA**

by

Alexander C. Nwala
M.S. May 2014, Old Dominion University
B.S. December 2011, Elizabeth City State University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2020

Approved by:

Michael L. Nelson (Director)

Michele C. Weigle (Member)

Jian Wu (Member)

Sampath Jayarathna (Member)

Ross Gore (Member)

ABSTRACT

BOOTSTRAPPING WEB ARCHIVE COLLECTIONS FROM MICRO-COLLECTIONS IN SOCIAL MEDIA

Alexander C. Nwala
Old Dominion University, 2020
Director: Dr. Michael L. Nelson

In a Web plagued by disappearing resources, Web archive collections provide a valuable means of preserving Web resources important to the study of past events. These archived collections start with seed URIs (Uniform Resource Identifiers) hand-selected by curators. Curators produce high quality seeds by removing non-relevant URIs and adding URIs from credible and authoritative sources, but this ability comes at a cost: it is time consuming to collect these seeds. The result of this is a shortage of curators, a lack of Web archive collections for various important news events, and a need for an automatic system for generating seeds.

We investigate the problem of generating seed URIs automatically, and explore the state of the art in collection building and seed selection. Attempts toward generating seeds automatically have mostly relied on scraping Web or social media Search Engine Result Pages (SERPs). In this work, we introduce a novel source for generating seeds from URIs in the threaded conversations of social media posts created by single or multiple users. Users on social media sites routinely create and share narratives about news events consisting of hand-selected URIs of news stories, tweets, videos, etc. In this work, we call these posts *Micro-collections*, whether shared on Reddit or Twitter, and we consider them as an important source for seeds. This is because, the effort taken to create Micro-collections is an indication of editorial activity and a demonstration of domain expertise. Therefore, we propose a model for generating seeds from Micro-collections. We begin by introducing a simple vocabulary, called *post class* for describing social media posts across different platforms, and extract seeds from the Micro-collections post class. We further propose Quality Proxies for seeds by extending the idea of collection comparison to evaluation, and present our Micro-collection/Quality Proxy (MCQP) framework for bootstrapping Web archive collections from Micro-collections in social media.

Copyright, 2020, by Alexander C. Nwala, All Rights Reserved.

Dedicated to my mother, Comfort C. Nwala and my father, Alexander E. Nwala.

ACKNOWLEDGMENTS

I am deeply grateful for God’s grace to come this far and the valuable contribution of so many that made this possible.

I am very grateful to my PhD supervisors, Dr. Michael L. Nelson and Dr. Michele C. Weigle. From the start of my PhD program in 2014 to the end in 2020, we had weekly meetings in which I received feedback and guidance. Their supervision covered the art and science of research, evaluation, writing, presentation, etc. I’m very grateful to them.

The Web Science and Digital Libraries (WS-DL) research group (<https://ws-dl.cs.odu.edu/>) at Old Dominion University provided a healthy, helpful, and friendly environment to conduct research. From Sawood Alam, I learned so much from software solutions and best practices, to deployment. From Mat Kelly, I learned the art of publication and how to navigate conferences. Shawn Jones made me appreciate collaboration more, through our work on @StormyArchives (<https://oduwsdl.github.io/dsa-puddles/> and @storygraphbot (<http://storygraph.cs.odu.edu>). Mohamed Aturban and I joined WS-DL at almost the same time, he sat next to me, and worked hard with me.

In 2014, Dr. Stephan Olariu supervised my Masters thesis (*Generating Combinatorial Objects - A New Perspective*). He encouraged me to find solutions to problems like enumerating the states of placing balls into bins and the dynamics of stem cells (<https://doi.org/10.1093/bioinformatics/btw528>).

In 2017, Dr. Robert Faris, (researcher, Berkman Klein Center at Harvard) supervised my *2016 US Elections Media Manipulation* research. Our collaboration led to @storygraphbot and influenced my PhD research. In 2016, I worked with Adam Ziegler, Anastasia Aizman, and other researchers at the Harvard Library Innovation Lab on the Local Memory Project (<http://www.localmemory.org/>, @localmem) which influenced my research. I am very grateful for these collaborations.

From 2012 – 2014, I was a Masters research student funded by Ajay Gupta. Without his support, my PhD would not have materialized. He did not know much about me, but gave me a chance. I am very grateful for his support. I had a lot to learn having just recently graduated from college and a new research assistant. I am very grateful to Dr. Kalpesh Padia who was always there to show me how to research.

In 2007, I arrived the US a teenager at Elizabeth City State University, North Carolina. I thank Antonio Rook for teaching me my first lesson in C++. I thank Dr. Jamiiru

Luttamaguzzi who encouraged me to pursue novel mathematical contributions. Thank you to Dr. Ellis Lawrence and Dr. Kuldeep Rawat for giving the opportunity to do UAV research and teach High School students C++. But most especially, I am thankful to my cousin Dr. Kingsley Nwala and his family for accommodating and supporting me during the course of my undergraduate study.

Thank you Pastor Tibi Peters and the Renaissance Assembly family. To my parents, Alexander E. Nwala and Comfort C. Nwala (late) I cannot say thank you enough. My mother wanted me to study in the US even when there was no money. But she would always pray. My father made multiple sacrifices to make it happen. To the rest of my family, Hope, Chima, Joy, Oluchi, Udo, Vicky, Pastor Victor, and Portia, you all made this possible. Thank you.

TABLE OF CONTENTS

	Page
LIST OF TABLES	xxi
LIST OF FIGURES	xxx
Chapter	
1. INTRODUCTION	1
1.1 WHY WE NEED WEB ARCHIVE COLLECTIONS	1
1.2 THE SEED SELECTION PROBLEM	5
1.3 AUTOMATING THE SEED GENERATION PROCESS	9
1.4 RESEARCH QUESTIONS	17
2. BACKGROUND	20
2.1 THE WEB	20
2.2 WEB CRAWLING	22
2.3 CRAWLING SEARCH ENGINES	25
2.4 WEB ARCHIVING	26
2.5 SOCIAL MEDIA TOOLS FOR COLLECTION BUILDING	29
2.6 CHAPTER SUMMARY	38
3. RELATED WORK	40
3.1 COLLECTION BUILDING	40
3.2 SEED SELECTION	46
3.3 COLLECTION EVALUATION	50
3.4 CHAPTER SUMMARY	53
4. SCRAPING SEEDS FROM SERPS	54
4.1 EXPERIMENT: REFINING NEWS STORIES ON SERPS	55
4.2 RESULTS	61
4.3 GENERATING SEEDS FROM SERPS, A RECOMMENDATION	73
4.4 CHAPTER SUMMARY	74
5. SCRAPING SEEDS FROM MICRO-COLLECTIONS IN SOCIAL MEDIA	75
5.1 POST CLASS: CLASSIFICATION SYSTEM FOR LABELING SOCIAL MEDIA POSTS	75
5.2 EXPERIMENT: CHARACTERIZING AND COMPARING SERP AND MICRO-COLLECTION SEEDS	79
5.3 EVALUATION: METRICS FOR CHARACTERIZING AND COMPARING SERP AND MICRO-COLLECTION SEEDS	82
5.4 RESULTS	85
5.5 GENERATING SEEDS FROM SOCIAL MEDIA, A RECOMMENDATION	91

5.6	CHAPTER SUMMARY	92
6.	COMPARING COLLECTIONS OF SEEDS.....	93
6.1	COLLECTION CHARACTERIZING SUITE (CCS)	93
6.2	COLLECTION CHARACTERIZATION AND COMPARISON	101
6.3	EVALUATION	103
6.4	RESULTS	105
6.5	CHAPTER SUMMARY	108
7.	QUANTIFYING THE QUALITY OF SEEDS: QUALITY PROXIES (QPS) FOR SEEDS.....	109
7.1	QUALITY PROXIES (QP) FOR SEEDS	109
7.2	POPULARITY SEED QUALITY PROXY.....	111
7.3	NON-POPULARITY SEED QUALITY PROXIES	116
7.4	ADDITIONAL QUALITY PROXIES: FLIPPING QUALITY PROXIES	123
7.5	THE SEED QUALITY PROXY MATRIX AND COMPARING SEEDS	124
7.6	CHAPTER SUMMARY	125
8.	EXPLORING COLLECTIONS WITH QUALITY PROXIES.....	126
8.1	THE 2020 CORONAVIRUS PANDEMIC	127
8.2	THE FLINT WATER CRISIS	132
8.3	HURRICANE HARVEY	133
8.4	CHAPTER SUMMARY	135
9.	A FRAMEWORK FOR BOOTSTRAPPING WEB ARCHIVE COLLECTIONS FROM MICRO-COLLECTIONS IN SOCIAL MEDIA.....	137
9.1	FRAMEWORK OVERVIEW	137
9.2	FRAMEWORK EVALUATION	139
9.3	EVALUATION RESULTS AND DISCUSSION	146
9.4	LIMITATIONS OF FRAMEWORK.....	170
9.5	GENERATING SEEDS WITH THE MCQP FRAMEWORK, A RECOMMENDATION	172
9.6	CHAPTER SUMMARY	175
10.	CONTRIBUTIONS, FUTURE WORK, AND CONCLUSIONS.....	177
10.1	CONTRIBUTIONS	179
10.2	FUTURE WORK	180
10.3	CONCLUSIONS	181
	REFERENCES.....	200
	APPENDICES	
A.	EVALUATION RESULTS: ADDITIONAL TABLES FOR TOP 10 OVERLAP (WITH P@10) FOR SEEDS SCORED BY 1 – 3 QP COMBINATIONS	201

B.	EVALUATION RESULTS: ADDITIONAL TABLES FOR AVERAGE OVERLAP AND AVERAGE P@K	207
C.	EVALUATION RESULTS: ADDITIONAL TABLES FOR AVERAGE P@K FOR ADDITIONAL OVERLAP INTERVALS OF QP COMBINATIONS ...	215
D.	EVALUATION RESULTS: SUPPLEMENTARY LINE VIS FOR OVERLAP AND P@K FOR DIFFERENT COMBINATIONS OF QUALITY PROXIES	222
E.	EVALUATION RESULTS: ADDITIONAL TABLES FOR AVERAGE OVERLAP AND AVERAGE P@10 FOR DIFFERENT COMBINATIONS OF QUALITY PROXIES	226
F.	EVALUATION RESULTS: EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION OF THE DIVERSITY (D_U - UNIQUE RATIO) OF REFERENCE AND MICRO-COLLECTION SEEDS	230
G.	EVALUATION RESULTS: EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION OF THE DIVERSITY (D_C - SIZE CHANGE AFTER COMPRESSION) OF REFERENCE AND MICRO-COLLECTION SEEDS	243
	VITA	256

LIST OF TABLES

Table		Page
1	A list of the first five URIs extracted from different collection building sources (Google and Twitter) and a Tweet reply thread [34], for the <i>Flint Water Crisis</i> story extracted 2018-11-07. The entries are sorted in reverse order of publication.	10
2	Sample of seed URLs from Archive-It <i>Ebola virus</i> collection, URLs extracted from Reddit SERP (Search Engine Result Page) and comments for query “Ebola virus,” and URLs extracted from the references of the Wikipedia <i>Ebola virus</i> document.	18
3	Gossen et al. [108]: Exemplary scopes used in a sub-collection specification. This list is not exhaustive.	42
4	Gossen et al. [110]: Examples of temporal event characteristics.	43
5	Summary of Previous Research and This Research ’s approach toward collection building.	49
6	The RLG Collecting Levels	52
7	The RLG Collecting Levels Language Suffixes	52
8	The <i>SERP-Refind</i> dataset [141] generated by extracting URIs from SERPs (<i>General</i> and <i>News</i> vertical) for seven queries between 2017-05-25 and 2018-01-12.	57
9	Average story replacement rate for <i>General</i> and <i>News</i> vertical SERP collections. Column markers: minimum⁻ and maximum⁺	62
10	Average new story rate for <i>General</i> and <i>News</i> vertical SERP collections. Column markers: minimum⁻ and maximum⁺	63
11	Probability of finding the same story after one day, one week, and one month (from first observation) for <i>General</i> and <i>News</i> vertical SERP collections. Column markers: minimum⁻ and maximum⁺	65

12	Comparison of two collections against the <i>June-2017</i> collection (documents published in June 2017). The collection <i>Jan-2018</i> , which was created (2018-01-11) without modifying the SERP date range parameter has a lower overlap than the collection (<i>June-2018-Restricted-to-June</i>) created the same day (2018-01-11) by setting the SERP date range parameter to June 2017. Even though setting the date range parameter increases finding stories with common publication dates as the date range, the recall is poor due to the fixed SERP result. Column markers: maximum	72
13	Post class for social media posts. All non- $\mathbf{P}_1\mathbf{A}_1$ collections are combined to create Micro-Collections (MC). However, some $\mathbf{P}_1\mathbf{A}_1$ posts (e.g., Figure 32) can be considered as Micro-collections if they contain more links than the median number of links estimated for $\mathbf{P}_1\mathbf{A}_1$ posts of the social media platform.	76
14	Temporal characteristics of the Micro-collections dataset topics	80
15	Post class counts (Class), Social media posts (Posts), and URI counts (URIs) for dataset generated by extracting URIs from post classes ($\mathbf{P}_1\mathbf{A}_1$, $\mathbf{P}_n\mathbf{A}_1$, and $\mathbf{P}_n\mathbf{A}_n$) of Reddit, Twitter, Twitter Moments, and Scoop.it. The Micro-collection (MC) post class is formed by combining posts in $\mathbf{P}_n\mathbf{A}_1$ and $\mathbf{P}_n\mathbf{A}_n$ post classes. .	81
16	Probability (e.g., $P(p_{\mathbf{P}_1\mathbf{A}_1}^{Reddit} = 1) = \mathbf{0.63}$) of the event that a social media post from a given post class (e.g., Reddit $\mathbf{P}_1\mathbf{A}_1$) has k HTML URIs (e.g., $k = 1$).....	85
17	Conditional probability (e.g., $P(relevant p_{\mathbf{P}_1\mathbf{A}_1}^{Reddit} = 1) = \mathbf{0.64}$) of the event that the URIs in a social media post from a given post class (e.g., Reddit $\mathbf{P}_1\mathbf{A}_1$) are relevant, given that the post has k (e.g., $k = 1$) HTML URIs. Column markers: minimum and maximum . For the $\mathbf{P}_1\mathbf{A}_1$, $k = 5+$ Twitter cell, the probability was calculated for just one post with eight HTML URIs.	86
18	Summary recommendations of the Source to prioritize when generating seeds from social media based on the Attribute Prioritized, Query Type, and Vertical	91
19	Distribution of Top Five Topics for Two Archive-It Collections.	96
20	foo	102
21	The <i>CCS</i> Evaluation Dataset comprised of 129 collections from three Topics: “Ebola Virus,” “Hurricane Harvey,” and “2016 Pulse Nightclub shooting.” WSDL represents the collections generated by the authors.	104
22	List of collections most similar to three Archive-It collections and three random collections for the evaluation dataset topics.	105

23	Ranking of CCS Metrics based on the Standard Deviation of CCS values in the dataset	108
24	Summary of the Quality Proxies (QPs) for seeds.	111
25	The ap_i values of four seeds from Figure 37. The <i>in</i> and <i>out</i> -degree details were extracted on February 15, 2020. The $offset = 0$ since the minimum d_i , $3,723 \geq 0$. The difference between the minimum and maximum d_i , $(5,394,414 - 5,398,137 - 3,723)$ was used to normalize ap_i . dp_i is calculated in the same fashion as ap_i with one important difference: the <i>in</i> and <i>out</i> -degree information is extracted from the Twitter handle that has a bi-directional (Figure 38) link with the domain of the seed.	115
26	Illustration of the assignment of <i>su</i> scores for two seed domains (<code>cdc.gov</code> vs. <code>espn.com</code>) for the query “ebola virus.”. We use the count of result pages from the Google SERP (Figure 40) to estimate the <i>subject expertise</i> of the domain of a seed. Accordingly, <code>cdc.gov</code> has a higher subject expertise than <code>espn.com</code> . These counts were derived from queries issued on February 16, 2020.	120
27	For <i>The 2020 Coronavirus Pandemic</i> , top five seeds extracted by combining three popularity-based QPs rp , sh , lk to produce a single QP score (q - Equation 13), ranking the seeds by their QP scores, and selecting the top five seeds with the highest scores. The table illustrates how popularity-based Quality Proxies unsurprisingly gives more credit to seeds from popular (well-known) domains.	127
28	For <i>The 2020 Coronavirus Pandemic</i> , top five seeds extracted by combining geographical QPs ge_a and ge_d , ranking the seeds by their QP scores (q - Equation 13), and selecting the top five seeds with the highest scores. The table illustrates the interplay of ge_a and ge_d by showing how authors from different geographical regions share seeds from different domains.	129
29	For <i>The 2020 Coronavirus Pandemic</i> , top five seeds with the highest <i>broad reputation</i> QP score (re_b). For a single seed (e.g., <code>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1592694/</code>), the re_b score (e.g., 0.81) was approximated by counting the number of times the seed domain (e.g., <code>.nih.gov</code>) was cited (e.g., 46 times) in a reputation gold standard of 57 representative Wikipedia documents (one vote per document) about <i>Disease outbreaks</i>	131
30	For <i>The Flint Water Crisis</i> , top five seeds for different domains extracted by combining relevance rl and the unflipped (ge_d) and flipped ($\overline{ge_d}$) geographical QP. ge_d (Section I) helped in surfacing local media (e.g., <code>mlive.com</code> and <code>detroitnews.com</code>) while $\overline{ge_d}$ rewarded seeds from news media organizations distant (e.g., <code>bbc.com</code> and <code>theguardian.com</code>) from Flint, Michigan.	133

- 31 For *The Flint Water Crisis*, top five seeds extracted by focusing on broad (referenced across *Public health crisis* topics - Section I, re_b) and narrow (referenced only in the *Flint Water Crisis* story - Section II, re_n) reputation. For re_b , **Hits** represents the count of Wikipedia documents (one document - one vote) that cite a domain from a gold standard collection of 70 *Public health crisis* Wikipedia reference documents. For re_n , it represents the number of times (out of 550 references) a domain was cited in the Wikipedia *Flint Water Crisis* document. Broadly-defined reputation benefits well-known (e.g., `nih.gov` and `nytimes.com`) organizations. Narrowly-defined reputation benefits local media (e.g., `mlive.com`, `abc12.com`, `freep.com`). 134
- 32 For *Hurricane Harvey*, top five seeds extracted by combining relevance (rl) and the unflipped scarcity Quality Proxy sc (Section I) and flipped \bar{sc} (Section II). Scarcity can be used to increase the diversity of the seed domains as reflected by the domains (e.g, `texasmonthly.com`, `eonline.com`, and `espn.com`), flipping the Quality Proxy results in surfacing seeds from domains (e.g., `cnn.com` and `abcnews.go.com`) that appear multiple times in the collection. 136
- 33 Framework evaluation dataset [191] consisting of 1,552 seeds from Reference (Google & Expert) collections, and 2,027 seeds from 4,209 tweets from Twitter Top/Latest Micro-collections extracted at different date ranges. 141
- 34 List of Quality Proxies extracted from evaluation dataset seeds. We additionally included the flipped states of these Quality Proxies for scoring seeds. 142
- 35 Precision gold-standard dataset. The documents from the references of these Wikipedia articles were used to generate document vectors for measuring relevance. Relevance was approximated by the similarity between a seed's document vector and the gold-standard vector corresponding to the seed's topic. Similarity exceeding the specified relevance threshold signaled the relevance of the seed. . . . 143
- 36 A sample of 12 QP combinatorial states for 1-combination, 2-combination, and 3-combinations. A single 1-combination or 2-combination or r -combination of QPs can be used to score (Equation 15) a seed. 144

- 37 (Chapter 9.3.1, Coronavirus): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP Combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 147
- 38 Top 10 seeds from Micro-collections extracted by two different QP combinations: rl, re_b (left) and $rl, \overline{lk}, \overline{ge}_a, rt$ (right). The QP scores prefix the domains. The left with 1.0 overlap with Google (precision: 0.6) consists of popular domains (e.g., `nytimes.com` and `who.int`) while the right (0.0 overlap, 0.6 precision) consists of less popular and international (due to \overline{ge}_a) domains (e.g., `bylinetimes.com` and `rappler.com`). Non-relevant seeds have been struck through. 148
- 39 (Chapter 9.3.1, Coronavirus, Supplements Table 37 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Figure 45 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - Expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of Expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 153
- 40 Precision for reference (Google - G, Expert - E) and Micro-collections (M) seeds. Reference collections G, E produced seeds of a higher precision than M. M seeds were all above the relevance threshold (Table 35) except M from *2018 World Cup* - Twitter-Latest (*). These values were populated from the last rows of all the tables in Appendix B and Table 39. 154
- 41 (Chapter 9.3.1 & 9.3.2, Coronavirus, Supplements Table 37 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average. 156

42 4. (Chapter 9.3.1, Coronavirus, Supplement Table 37 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row). 164

43 Median of the average overlap (\max^+ , \min^-) and P@10 for dataset topics for lower order (1 – 3) and higher order (4 – 10 and All) combinations. The combination *All* means that all Quality Proxies (without flipped state) where used to score the seeds. 166

44 First (**Q1**), second (**Q2**), and third (**Q3**) quartiles of Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of reference (Google - **G** & Expert - **E**) and Micro-collection (M_E & M_G) seeds. A single cell, e.g., **G**, **Q1**, from **No. 1**, reads as follows 25% of seeds had diversity \leq **0.60**. Overall, seeds (with r -superscript) selected without using QP scores has a higher diversity of seeds selected with QP scores, Google seeds had the highest diversity, while the diversity of Micro-collection and Experts seeds was similar. **Key:** **green** - column-wise maximum, **red** - column-wise minimum. 168

45 First (**Q1**), second (**Q2**), and third (**Q3**) quartiles of Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of reference (Google - **G** & Expert - **E**) and Micro-collection (M_E & M_G) seeds. A single cell, e.g., **G**, **Q1**, from **No. 1**, reads as follows 25% of seeds had diversity \leq **0.60**. Overall, seeds (with r -superscript) selected without using QP scores has a higher diversity of seeds selected with QP scores, Google seeds had the highest diversity, while the diversity of Micro-collection and Experts seeds was similar. **Key:** **green** - column-wise maximum, **red** - column-wise minimum. 169

46 (Chapter 9.3.1, 2018 World Cup): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 202

- 47 (Chapter 9.3.1, Hurricane Harvey (collected 2020)): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 203
- 48 (Chapter 9.3.1, Hurricane Harvey (collected 2017)): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 204
- 49 (Chapter 9.3.1, Flint Water Crisis): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 205

- 50 (Chapter 9.3.1, 2014 Ebola Virus Outbreak): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, reb), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 206
- 51 (Chapter 9.3.1, Coronavirus-Latest, Variant of Table 39 for Twitter-Latest seeds): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 45 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 208
- 52 (Chapter 9.3.1, 2018 World Cup, Supplements Table 46 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 46 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 209

- 53 (Chapter 9.3.1, 2018 World Cup-Latest, Variant of Table 52 for Twitter-Latest seeds): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 46 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 210
- 54 (Chapter 9.3.1, Hurricane Harvey (collected 2020), Supplements Table 47 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 47 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 211
- 55 (Chapter 9.3.1, Hurricane Harvey (collected 2017), Supplements Table 48 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 47 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 212

- 56 (Chapter 9.3.1, Flint Water Crisis, Supplements Table 49 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 48 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 213
- 57 (Chapter 9.3.1, 2014 Ebola Virus Outbreak, Supplements Table 50 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 48 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r). 214
- 58 (Chapter 9.3.1 & 9.3.2, Coronavirus-Expert reference (E), Supplements Table 37 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average. 216
- 59 (Chapter 9.3.1 & 9.3.2, 2018 World Cup, Supplements Table 46 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average. 217
- 60 (Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2020), Supplements Table 47 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average. 218

61	(Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2017), Supplements Table 48 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.	219
62	(Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2017)-Expert reference (E), Supplements Table 48 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (left) of QP combinations of different overlap ranges and the count of QP combinations (right) that produced the average.....	219
63	(Chapter 9.3.1 & 9.3.2, Flint water crisis, Supplements Table 49 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.	220
64	(Chapter 9.3.1 & 9.3.2, 2014 Ebola Virus Outbreak, Supplements Table 50 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (left) of QP combinations of different overlap ranges and the count of QP combinations (right) that produced the average.	221
65	(Chapter 9.3.1 & 9.3.2, 2014 Ebola virus outbreak-Expert reference (E), Supplements Table 50 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.....	221
66	4. (Chapter 9.3.1, 2018 World Cup, Supplement Table 46 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).	226
67	4. (Chapter 9.3.1, Hurricane Harvey (collected 2020), Supplement Table 47 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).....	227
68	4. (Chapter 9.3.1, Hurricane Harvey (collected 2017), Supplement Table 48 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).....	228

69	4. (Chapter 9.3.1, Flint water crisis, Supplement Table 49 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).	229
70	4. (Chapter 9.3.1, 2014 Ebola Virus Outbreak, Supplement Table 50 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).	229

LIST OF FIGURES

Figure	Page
1 The NLM Archive-It <i>Ebola Virus</i> Collection [3] showing the Seed URLs.....	3
2 A USAID Webpage [19] from the NLM Archive-It <i>Ebola Virus</i> Seeds	4
3 A CDC Webpage [20] from the NLM Archive-It <i>Ebola Virus</i> Seeds	4
4 The Internet Archive has on multiple occasions requested that users submit seeds to bootstrap collections. The time when users respond with seeds impacts the collections generated.	7
5 The Wikipedia page [35] for the MSD High School shooting, created the same day as the shooting event (February 14, 2018).	12
6 References from Wikipedia MSD High School shooting page. As of July 2, 2020, it had 271 references. We propose extracting URIs from Micro-collections such as this to generate seeds.	12
7 A Twitter Moment [37] about the Stoneman Douglas High School shooting created the day after (February 15, 2018) the tragic incident. Social media Micro-collections such as this provides the opportunity for creating seeds to bootstrap archived collections. This is especially useful when no archived collection for the event exist; as of July 2, 2020, there was no Archive-It collection for the <i>Stoneman Douglas High School</i> shooting event. This screenshot has been edited to show more detail.	14
8 The Micro-collection from Storify (a) for the <i>Ukrainian crisis</i> event was created in January 2014 and highlights incidents such as riots before the event became a prominent news event. Russia began the annexation of Crimea in late February coinciding with the creation of the Archive-It collection (b). The Archive-It collection potentially omits some of the prelude contents in the Storify Micro-collection (a).	16
9 W3C [43]: Illustration showing the relationship between URI, Resource, and Representation.	21
10 Illustration showing the Web Crawling Process.	23
11 Illustration showing the Focused Crawling Process. The blue annotation marks the new parts added to a Web crawler to convert it to a Focused Crawler.	24

12	Targets of Web and Focused Crawlers. Web crawlers used by search engines build (and update) indexes without taking the topics of documents into consideration, but focused crawlers focus on collecting documents that are similar to a narrow set of topics.	25
13	Helen Hockx-Yu [80]: Key Processes of Web Archiving.	27
14	Memento Framework [86]: Architectural overview of how the Memento framework allows accessing a prior version of a resource.	28
15	A pair of social media posts consisting of multiple hand-selected URIs. The authors of posts such as these may not consider their posts as Micro-collections or the URIs as seeds, however, these posts exemplify collection building activity.	30
16	<i>Nice Threads</i> , a feature for creating a collection of tweets by clicking the plus sign.	32
17	A tweet reply thread [95] about the <i>Flint Water Crisis</i> from Senator Tammy Duckworth consisting exclusively of text (no URIs or images). Reply threads are formed by replying to each preceding tweet, and thus provides an implicit means of creating a collection of tweets.	33
18	A tweet reply thread [96] about the <i>Flint Water Crisis</i> from <i>March for Science</i> includes a single URI. The URIs embedded in reply threads may serve as seeds. .	34
19	A pair of three tweets that are part of a reply thread [34] from <i>Doing Things Differently</i> about the <i>Flint Water Crisis</i> . This reply thread spans over 2.5 years, and consists of 74 tweets (as of October 29, 2018) each containing a URI.	35
20	A Facebook public post from <i>Asemeyibo Buowari-Brown</i> about the <i>StopTheSoot</i> movement. The post links to a 14-minute video about illegal refineries operating in the Niger-Delta of Nigeria. Such refineries contribute to the air pollution in the region. This post has been edited to show more content.	37
21	A Reddit post [101] from <i>jazir5</i> about Ebola virus vaccines. This post links to five authoritative sources that discuss the promise of immunity provided by Ebola virus vaccines.	38
22	Nanni et al. [112]: Overview of the method to extract event-centric sub-collections from Web archives	47
23	Google <i>General</i> (a) and <i>News</i> vertical (b) SERPs for the query “hurricane harvey.” Some links have been removed to enable showing more detail. For our experiment, links were extracted from the first five pages (annotation B) of both SERPs for each query.	55

24	A screenshot of the Google CAPTCHA page for query “bidden polls,” triggered by searching 18 times (paginations counted), each time paginating to a maximum of page 20.....	58
25	a & b: Page-level new story rates for <i>General</i> and <i>News</i> vertical SERPs. c & d: Page-level story replacement rates for <i>General</i> and <i>News</i> vertical SERPs.	64
26	a & b: Page-level probability of finding the URI of a story over time.	66
27	Probability of finding an arbitrary story for <i>General</i> and <i>News</i> vertical SERPs was modeled with two best-fit exponential functions. In general, the probability of finding the URI of a news story on the <i>General</i> SERP is higher (lower new story rate) than the probability of finding the same URI on the <i>News</i> vertical SERP (due to its higher new story rate).	67
28	Temporal distributions: Stories in <i>General</i> SERP collections (a & c) persist longer (“longer life”) than stories in <i>News</i> vertical collections (b & d). Compared to the “trump russia” <i>General</i> SERP collection, the stories in the “hurricane harvey” <i>News</i> vertical collection have a “longer life” due to a lower rate of new stories.	68
29	Page-level temporal distribution of stories in the “manchester bombing” <i>General</i> SERP collection showing multiple page movement patterns. Stories in <i>General</i> SERP collections persist longer than stories in <i>News</i> vertical collections. Color codes - page 1 , page 2 , page 3 , page 4, page 5 , and blank for outside pages 1 – 5.	70
30	Page-level temporal distribution of stories in the “manchester bombing” <i>News</i> vertical SERP collection showing multiple page movement patterns, and the shorter lifespan of <i>News</i> vertical URIs (compared to <i>General</i> SERP URIs). Color codes - page 1 , page 2 , page 3 , page 4, page 5 , and blank for outside pages 1 – 5.	71
31	Example of a Micro-collection from Twitter by a single author (@ScottGottliebMD) consisting of three tweets that are part of a reply thread [146] about the <i>2020 Coronavirus Pandemic</i> . This Micro-collection is of post class $\mathbf{P}_n\mathbf{A}_1$ since it consists of multiple Posts from a single Author. This image has been edited to show more details.....	77
32	Example of a pair of Micro-collection Reddit posts [147, 148] consisting of 102 external references for the 2014 Ebola outbreak. Both Micro-collections are of type $\mathbf{P}_1\mathbf{A}_1$ (single Posts from a single Author).	78
33	Ebola Virus Outbreak Precision Distribution: $\mathbf{P}_1\mathbf{A}_1$ seeds produced webpages with a higher precision than $\mathbf{P}_n\mathbf{A}_n$ for text but not hashtag queries. The black line marks the relevance threshold.	88

34	Ebola Virus Outbreak Age Distribution: MCs produced older webpages in the Twitter-Latest vertical for the older topics.	89
35	Distribution of CCS Metrics for pair of collections most similar to Archive-It collections (a – c).	107
36	Five-dimensional vector expressing the popularity of a seed embedded in a social media post.	112
37	Population of the <i>post popularity</i> dimensions of four seed popularity vectors from the <i>replies</i> , <i>likes</i> , and <i>shares</i> statistics of their respective containing tweets. The <i>post popularity</i> dimensions can be additionally populated with k social media posts that embed the seed.	113
38	An illustration of a bi-directional link; the Twitter account @WHO (left) links to the <code>who.int</code> front page (right), and the <code>who.int</code> front page (right) links to the @WHO (left) Twitter account. The presence of a bi-directional link validates that the <code>who.int</code> domain is associated with @WHO Twitter handle, and the use of @WHO as a source for the <i>in</i> and <i>out</i> -degree information needed to calculate dp . The screenshot on the right has been edited to show more detail.	116
39	Seven additional non-popularity (proximity and uncategorized) dimensions of the seed authority vector expressing quality of the seed across <i>geographical</i> (author and domain), <i>temporal</i> , <i>subject expert</i> , <i>retrievability</i> , <i>relevance</i> , <i>reputation</i> (broad and narrow), and <i>scarcity</i> dimensions.	117
40	Google SERP showing (red annotation) the count (951,000) of result pages for query: “site:cdc.gov” - an estimate of the total number of pages from <code>cdc.gov</code> indexed by Google. We use this statistic from the SERP to calculate (Table 26) the <i>su</i> score for a seed.	121
41	A seed Quality Proxy matrix \mathbf{Q} . Each row represents a 14-dimensional seed Quality Proxy vector \mathbf{q} for a seed $seed_i$	124
42	The seed: <code>https://jesusislordradio.info/</code> (Table 28, Section II, No. 5) shared by @_lameckonger from Kisii, Kenya, was surfaced by prioritizing authors ($\overline{ge_a}$) and domains ($\overline{ge_d}$) distant from New York City. This seed could be considered non-relevant, however, if the context of concern requires supplying domains that satisfy the condition <i>religious responses to the Coronavirus Pandemic</i> , the seed could be considered relevant.	130
43	MCQP framework overview for bootstrapping Web archive collections from Micro-collections in Social Media. The numbers shown represent the stages of the framework.	138

- 44 Overlap vs P@20 for Google (G - orange dots) and Micro-collection (M - blue dots) *2020 Coronavirus Pandemic* Twitter-Latest seeds scored by different Quality Proxies. A single dot represents the overlap (X-axis) and P@20 (Y-axis) for seeds scored by a single Quality Proxy. The scatterplot shows how different Quality Proxy scores result in high (e.g., dp, ge_a, re_b and ge_d, re_b, re_n) or low (rl, ge_d, re_n) overlap/P@20. Unsurprisingly, the QP combination rl, ge_d, re_n resulted in a low P@20 because the *relevance rl* QP was flipped, meaning relevance was penalized. 150
- 45 (Chapter 9.3.1, *Coronavirus, Supplements Table 39 & 51*): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For Twitter-Top (first column), as K increased and overlap dropped, the average P@K for M_G and M_E (solid lines) mostly held steadily, with median of 0.71 ($\sigma = 0.05$) and 0.70 ($\sigma = 0.08$), respectively, a 0.16 (M_G) and 0.15 (M_E) increase above the baseline (did not use QP scores) precision (0.55). Similarly, for Twitter-Latest (second column), as K increased, the average P@K for M_G and M_E (solid lines) mostly held steadily, with median of 0.71 ($\sigma = 0.03$) and 0.75 ($\sigma = 0.05$), respectively, a 0.16 (M_G) and 0.20 (M_E) increase above the baseline precision (0.55) which did not use QP scores. 157
- 46 (Chapter 9.3.1, *2018 World Cup, Supplements Table 52 & 53*): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For Twitter-Top (first column), as K increased and overlap dropped, the average P@K for M_G (solid line) mostly held steadily, with median of 0.53 ($\sigma = 0.11$), a 0.09 increase above the baseline (did not use QP scores) precision (0.44). In contrast, for Twitter-Latest (second column), the utilization of QP scores did not improve the median P@K as K increased which might be attributed to the fact that the seeds came from a collection with the second lowest median average P@K, and thus, the QP scores could not improve already poor-performing seeds. 158
- 47 (Chapter 9.3.1, *Hurricane Harvey (collected 2020/2017), Supplements Table 54 & 55*): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For the collection collected in 2020 (first column), three years after the event, the utilization of QP scores did not improve the median P@K as K increased, which might be attributed to the fact that the seeds came from a collection with the lowest median average P@K, and thus, the QP scores could not improve already poor-performing seeds. In contrast, for the collection collected in 2017 (second column), as K increased and overlap dropped, the average P@K for M_G (solid lines) mostly held steadily, with median of 0.34 ($\sigma = 0.14$), a 0.19 increase above the baseline (did not use QP scores) precision (0.15). However, unlike M_G , the utilization of QP scores did not improve M_E 159

- 48 (Chapter 9.3.1, Flint Water Crisis & 2014 Ebola Virus Outbreak, Supplements Table 56 & 57): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For *Flint Water Crisis* (first column), as K increased and overlap dropped, the average P@K for M_G (solid line) mostly held steadily, with median of 0.60 ($\sigma = 0.15$), a 0.15 increase above the baseline (did not use QP scores) precision (0.45). For *2014 Ebola Virus Outbreak* (second column), with and without the use of QP scores led to a marginal improvement (≤ 0.01) above the baseline precision (0.24)..... 160
- 49 (Chapter 9.3.1, Coronavirus & 2018 World Cup, Supplements Table 41 (Coronavirus) and Table 59 (2018 World Cup) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., row 1, Coronavirus - G/M, overlap of 0; P@K = 0.70/0.48) or high overlap (e.g., Coronavirus, overlap (0.70, 0.80], P@K = 0.62/0.59) between K top seeds (scored by QP scores). The black line (0.20) marks the relevance threshold. . . . 161
- 50 (Chapter 9.3.1, Hurricane Harvey (collected 2020 and 2017), Supplements Table 60 (collected 2017) and Table 61 (collected 2017) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., first row - G/M, overlap of 0; P@K = 0.42/0.08) or high overlap (e.g., second row, overlap 0.80 – 0.90, P@K = 0.84/0.57) between K top seeds (scored by QP scores). Intuitively the higher overlap between reference seeds (high quality) and Micro-collection, the higher P@K for Micro-collections. The first and second charts aligns the most with this intuition unlike the third. The black line (0.10) marks the relevance threshold. 162
- 51 (Chapter 9.3.1, Flint Water Crisis and 2014 Ebola Virus Outbreak, Supplements Table 64 (Flint Water Crisis) and Table 63 (2014 Ebola Virus Outbreak) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., Flint Water Crisis - G/M, overlap of 0; P@K = 0.80/0.46) or high overlap (e.g., Flint Water Crisis, overlap 0.40 - 0.50, P@K = 0.93/0.81) between K top seeds (scored by QP scores). The black line (0.20) marks the relevance threshold. 163
- 52 (Chapter 9.3.1, Coronavirus and 2018 World Cup, Supplementary Line chart visualization of Table 42 (Coronavirus) and Table 66): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies 223

53	(Chapter 9.3.1, Hurricane Harvey (collected 2020 and 2017), Supplementary Line chart visualization of Table 67 (collected 2020) and Table 68 (collected 2017)): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies	224
54	(Chapter 9.3.1, Flint Water Crisis and 2014 Ebola Virus Outbreak, Supplementary Line chart visualization of Table 69 (Flint Water Crisis) and Table 70 (2014 Ebola Virus Outbreak)): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies	225
55	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	231
56	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	232
57	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	233
58	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	234
59	(Chapter 9.3.4, 2018 World Cup-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	235
60	(Chapter 9.3.4, 2018 World Cup-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	236
61	(Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	237

62	(Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	238
63	(Chapter 9.3.4, Hurricane Harvey (collected 2020), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	239
64	(Chapter 9.3.4, Flint Water Crisis, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.....	240
65	(Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	241
66	(Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	242
67	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	244
68	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	245
69	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	246
70	(Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	247

71	(Chapter 9.3.4, 2018 World Cup-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	248
72	(Chapter 9.3.4, 2018 World Cup-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	249
73	(Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	250
74	(Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	251
75	(Chapter 9.3.4, Hurricane Harvey (collected 2020), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	252
76	(Chapter 9.3.4, Flint Water Crisis, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	253
77	(Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	254
78	(Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.	255

CHAPTER 1

INTRODUCTION

On March 23, 2014, the World Health Organization reported the largest outbreak of Ebola virus in history in the forested region of rural southeastern Guinea [1]. After the initial reports, similar outbreaks were reported in neighboring Western African countries of Liberia and Sierra Leone. From the initial reports of Ebola in March 2014 to June 2016 [2], when the outbreak was declared over, Ebola had claimed the lives of over 11,000 people [1]. Two months after the Ebola outbreak was declared a Public Health Emergency of International Concern (PHEIC), an archivist at the National Library of Medicine (NLM) started collecting URLs to create a *Web archive collection* for the *Ebola virus* event [3]. A Web archive collection consists of groups of webpages that share a common topic e.g., “Ebola virus” or “Arab Spring.” The NLM *Ebola virus* Web archive collection includes websites of organizations, journalists, healthcare workers, and scientists, related to the 2014 Ebola virus discourse (Figures 1, 2, and 3). Collections such as the NLM *Ebola virus* collection are crucial to retrospective studies since they serve as time capsules that preserve the historic record of important events. This research explores the difficulties and ways of automatically generating the URLs (seeds) needed to build Web archive collections.

1.1 WHY WE NEED WEB ARCHIVE COLLECTIONS

In addition to the numerous and varied services the Web provides, such as social media and weather reporting, it is often the first place we go to learn about news stories and events. For example, on December 17, 2010, Mohamed Bouazizi, a fruit vendor in Tunisia, doused himself with paint thinner and set himself on fire outside the local government office [4]. His desperate act was to protest being arrested and beaten by local authorities for not having a permit to run a vegetable stall. His death ignited public anger giving way to street protests throughout Tunisia. The protests ultimately led to the end of the 23-year autocratic rule of President Zine el-Abidine Ben Ali [5]. The protests in Tunisia started a chain reaction of popular uprisings and protests in other Arab countries such as Egypt, Morocco, and Libya. These events are collectively described as the *Arab Spring* [6, 7, 8, 9, 10].

Today, 10 years after the death of Mohamed Bouazizi, Tunisia is a democratic republic with a functioning multi-party system. The Web houses many websites and social media

content that chronicle the Arab Spring. Unfortunately, as shown by SalahEldeen and Nelson, 11% of Web resources shared on social media are lost after the first year of publication [11]. This finding is not unique, there are many studies that show the decay of Web resources due to the problem of *link rot* and *content drift* [12, 13, 14, 15, 16]. Anyone who has ever clicked a link and was presented with a disappointing 404 response, indicating the absence of a resource, understands the impermanence of Web resources. Addressing this problem is critical since the Web holds a significant amount of our digital heritage. Fortunately, the link rot problem can and is being reduced through *Web archiving*, a process that involves collecting and persistently saving webpages in a digital archive. The Internet Archive¹ (IA), an organization founded in 1996, has been collecting and saving public webpages since its inception. This is based on a simple idea: an archived copy of a webpage may be viewed in place of a lost original copy, but this is only possible if the original webpage was saved. The ability to replay older versions of a webpage due to Web archiving has had significant implications including when there is a disagreement on “what was said.” For example, archived copies of webpages have been admitted as evidence in court cases [17, 18] and archived versions of social media content have been used to challenge politicians’ recollection of “what was said.” However, conventional Web archiving initiatives target general-purpose webpages, and while these are important, they are not well suited for stories and events that have a narrower scope. In other words, in addition to the preservation of general-purpose webpages, it is also important to preserve collections of webpages addressing a common topic such as the protests of the Arab Spring. This is the purpose of Web archive collections.

1.1.1 WEB ARCHIVE COLLECTIONS BEGIN AS SEEDS

A seed list (or seeds, or seed URLs) is an initial collection of the URLs of exemplar webpages for a topic. For example, Figure 1 contains a sample of four seed URLs (out of 144) from the NLM *Ebola virus* collection. The seed URLs and the pages they link to form a Web archive collection when crawled. *Crawling* is the process of discovering and saving URLs by visiting links which originate from a parent webpage, and subsequently visiting the links of its children pages and their respective descendants. This enables the discovery of the URLs of webpages that are linked to (directly and indirectly) from the parent page. Web archive collections begin with seeds, and quality seeds lead to quality Web archive collections. A collection of seeds having topics with variations of “buy cheap Rolex watches” or “we sell gold nuggets” is not expected to yield a good Web archive

¹<https://archive.org/>

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Group Sort By: **Count** | (A-Z)

Ebola Outbreak 2014 (144)
 Global Health Organizations (19)
 Measles (1)
 Nepal Earthquake 2015 (5)
 Zika Virus (34)

Language Sort By: **Count** | (A-Z)

English (119)
 Spanish (5)
 French (3)

Collector Sort By: **Count** | (A-Z)

National Library of Medicine (U.S.) (194)

Enter search terms here **Search** **Clear**

Sites Search Page Text

Page 1 of 3 (204 Total Results) **Next Page** ▶

Sort By: **Title (A-Z)** | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: WHO | Ebola and Marburg virus disease
URL: http://apps.who.int/iris/bitstream/10665/130160/1/WHO_HSE_PED_CED_2014.05_eng.pdf
 Captured once on Oct 17, 2014
 Group: Ebola Outbreak 2014
 Collector: National Library of Medicine (U.S.)

Title: Sierra Leone Ebola outbreak death toll now 5
URL: <http://bigstory.ap.org/article/sierra-leone-ebola-outbreak-death-toll-now-5>
 Captured once on Oct 28, 2014
 Group: Ebola Outbreak 2014
 Collector: National Library of Medicine (U.S.)

Title: On the Front Lines of an Epidemic: The Battle Against Ebola | USAID Impact
URL: <http://blog.usaid.gov/ebola/>
 Captured 47 times between Oct 22, 2014 and Apr 20, 2016
 Group: Ebola Outbreak 2014
 Language: English
 Collector: National Library of Medicine (U.S.)

Title: CDC - Blogs - CDC Director Blog
URL: <http://blogs.cdc.gov/cdcdirector/>
 Captured 42 times between Oct 15, 2014 and Mar 20, 2017
 Videos: 27 Videos Captured
 Group: Global Health Organizations
 Language: English
 Collector: National Library of Medicine (U.S.)

Fig. 1: The NLM Archive-It *Ebola Virus* Collection [3] showing the Seed URLs

You are viewing an archived web page, collected at the request of [National Library of Medicine](#) using [Archive-It](#). This page was captured on 9:32:44 Oct 22, 2014, and is part of the [Global Health Events web archive](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.



USAID
FROM THE AMERICAN PEOPLE

IMPACT

USAID HOME ABOUT THIS BLOG ARCHIVES RSS Feed

Search for:

ON THE FRONT LINES OF EBOLA

ON THE FRONT LINES OF AN EPIDEMIC: THE BATTLE AGAINST EBOLA

BLOG ARCHIVES
Select Month:

TAG CLOUD

16 Days Afghanistan Africa Agriculture AIDS Child survival Climate Change Democracy Development Disaster Assistance Education

Today the world is facing the largest and most-protracted Ebola epidemic in history and President Obama has declared it a top national security priority.

"Faced with this outbreak, the world is looking to us, the United States, and it's a responsibility that we embrace. We're prepared to take leadership on this to provide the kinds of capabilities that only America has, and to mobilize the world in ways that only America can do. That's what we're doing as we speak."

Fig. 2: A USAID Webpage [19] from the NLM Archive-It *Ebola Virus Seeds*

You are viewing an archived web page, collected at the request of [National Library of Medicine](#) using [Archive-It](#). This page was captured on 9:35:01 Oct 15, 2014, and is part of the [Global Health Events web archive](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.

CDC Home
CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

A-Z Index: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

CDC Director Blog
Thoughts from CDC Director Tom Frieden, MD, MPH

Guest Blog: Calling all Innovators to Help Fight Ebola

Categories: Health Protection

October 9th, 2014 3:39 pm ET -
Posted by [Dr. Rajiv Shah](#) on Thursday, October 9th 2014

Ed. note: This is cross-posted on "USAID From the American People Blog." See the original post [here](#).

Saving lives at birth. Powering clean energy solutions in agriculture. Inventing new tools to teach a child to read. Across development, we're calling on the world's brightest minds to tackle our toughest challenges. In the last few years, we have helped launch five Grand Challenges for Development that have rallied students and scientists, innovators and entrepreneurs to tackle some of humanity's toughest problems.

Today, we face just that kind of challenge—a global health crisis that is in dire need of new ideas and bold solutions. From Guinea to Liberia to Sierra Leone, Ebola is devastating thousands of families, disrupting growth, and fraying the fabric of society. The United States is helping lead the global response to the epidemic, but we cannot do it alone. That is why President Obama launched our sixth Grand Challenge, *Fighting Ebola: A Grand Challenge for Development* is designed provide health care workers on the front lines with better tools to battle Ebola.

- [Read more](#)

Guest Blog: The President Meets with Senior Staff to Discuss the U.S. Response to Ebola

Categories: Health Protection

October 9th, 2014 3:29 pm ET -

By: [David Hudson](#), Associate Director of Content for the Office of Digital Strategy, White House

Ed. note: This is cross-posted on "The White House Blog." See the original post [here](#).

This afternoon, President Obama met with his senior health, homeland security, and national security advisors to review the United States' response to the Ebola epidemic.

Blog Categories

- Director's Briefing Videos
- Health Protection
- Leading Causes of Death, Disability, & Injury
- Public Health & Clinical Care Collaboration

About this Blog

- About Dr. Frieden
- Guidelines for Public Comments

Contact Us:

Centers for Disease Control and Prevention

Fig. 3: A CDC Webpage [20] from the NLM Archive-It *Ebola Virus Seeds*

collection for the “2009 Swine Flu outbreak.” This is where curators come into the picture. Curators must ensure they select seeds that are relevant to the collection topic. This means curators (such as the NLM archivists) not only have the responsibility of searching for URLs to populate the seed list, but they also serve as filters to remove non-relevant URLs. This is a time-consuming process because it is mostly done manually.

1.2 THE SEED SELECTION PROBLEM

Web archive collections preserve Web resources that are relevant to specific topics ranging from disease outbreaks to popular uprisings, and provide the means to go back in time to study events which may no longer be effectively represented on the live Web due to link rot. These collections begin with seeds often selected manually by a curator. Curator-generated seeds have some important advantages. First, seeds selected by expert (human) curators are often of a high quality since the curators can easily identify and remove off-topic URLs. Second, curators have the ability to create a collection tailored to the specific needs of the collection topic, whether narrow or broad. For example, a health expert could build three narrowly scoped collections for the 2014 Ebola outbreak, the 2016 outbreak, and the 2018 outbreak. This flexibility is hard to automate, as this work will show, but the curator method of manually selecting URLs as seeds is limited in multiple ways. First, it is time consuming to collect seeds. For example, it took several months to collect the NLM *Ebola virus* seeds. Second and most crucial, the curator method of generating seeds requires domain knowledge of the collection topic. For example, building a collection about *the local government response after the death of Mohamed Bouazizi* requires domain knowledge of the socio-political and cultural environment in Tunisia. The average American citizen does not have this knowledge, and thus cannot effectively collect seeds for this event. Collecting seeds for this event is made even more difficult given the fact that some relevant seeds may be in the Arabic language. Unfortunately, there is a shortage of curators to collect seeds for rapidly unfolding local and global events, so we cannot rely exclusively on human curators such as the NLM archivists or curators at the Archive-It to build Web archive collections. To cope with this shortage of curators amidst an abundance of world events, various organizations such as the Internet Archive routinely request (Figure 4) for users to contribute links (seeds) for Archive-It collections, e.g., the *2016 Pulse Nightclub Shooting* [21], the *2016 U.S. Presidential Election* [22], and the *Dakota Access Pipeline* [23] collections. But this crowdsourced approach to collection building is not enough, because it means letting Archive-It or similar collection building organizations decide what events or stories

are important and worth building Web archive collections for. Besides, even when there is consensus on what events are worth building Web archive collections for, the experts might be unavailable. In some other cases, the collections are initiated months or years after the precipitating event. This could have serious consequences especially for long-running events: Web archive collections that start late could omit webpages that address the early stages of events [24, 25]. The omission is analogous to writing a story book with the first few chapters missing. Let us consider two important stories which illustrate this omission and the absence of Web archive collections for important events.

On February 14, 2018, there was a tragic shooting that claimed the lives of 17 people at the Marjory Stoneman Douglas (MSD) High School in Florida. In the aftermath of the tragic event, the teenage students boldly stepped into the highly politically divisive gun control debate demanding stricter gun control measures [26, 27]. Less than two weeks after the shooting, major gun sellers Walmart and Dick's Sporting increased the minimum age required to purchase firearms and ammunition from 18 to 21 [28, 29]. Dick's additionally discontinued the sale of assault-style rifles, high capacity magazines, and bump stocks. On March 9, 2018, Governor Rick Scott of Florida signed the *Marjory Stoneman Douglas High School Public Safety Act* bill into law. Among other gun control measures, it raised the minimum age for buying rifles to 21, banned bump stocks, and instituted background checks. The Stoneman Douglas students refocused national attention on the gun control problem by holding the *March for Our Lives* demonstration in Washington, DC on March 24, 2018, with at least 1.2 million people in attendance [30]. It was reported as the biggest youth protest since the Vietnam War [30]. The ripple effects of the activism of the Stoneman Douglas students is still being felt, and most would agree that this incident deserves highlight as part of the broader gun control discourse in the United States, thus worthy of a Web archive collection. But as of July 2, 2020, two years later, there was no Archive-It collection for the Stoneman Douglas shooting incident. It is fair to conjecture that we have already begun to lose social media and news content due to link rot. If we started building today, we would surely miss resources present two years ago. Consequently, it is important to start collecting seeds for Web archive collections early. This calls for a method for generating seeds automatically and on demand.

In April 2014, state officials in Flint, Michigan switched the city's water source from Lake Huron of the Detroit water system to the Flint River [31]. A month after the switch, Flint city residents complained about the water's taste and smell [32]. Between August and September 2014, the city issued three boil advisories to residents [31] after finding Fecal



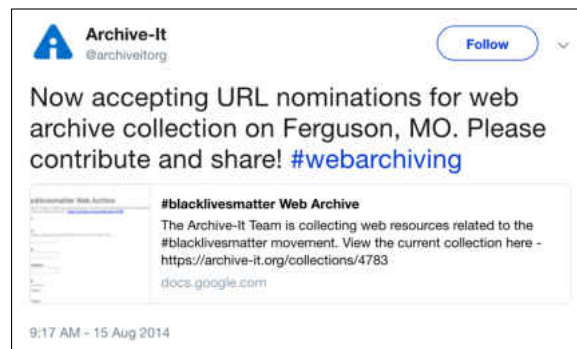
(a) A tweet from Archive-It requesting seeds for the 2012 *Hurricane Sandy* collection.



(b) A tweet from Archive-It requesting seeds for the 2013 *Boston Marathon Bombing* collection.



(c) A tweet from Archive-It requesting seeds for the 2013 *Nelson Mandela* collection.



(d) A tweet from Archive-It requesting seeds for the 2014 *Ferguson (#blacklivesmatter)* collection.



(e) A tweet from Archive-It requesting seeds for the 2014 *Ebola Virus* collection.



(f) A tweet from Archive-It requesting seeds for the 2016 *US Presidential Elections* collection.

Fig. 4: The Internet Archive has on multiple occasions requested that users submit seeds to bootstrap collections. The time when users respond with seeds impacts the collections generated.

Coliform Bacteria (E. Coli) in the water. Following multiple health incidents, on January 5, 2016, Michigan Governor Rick Snyder declared a state of emergency for the city of Flint, due to dangerously high levels of lead contamination in the drinking water. The flint story is known commonly today as the *Flint Water Crisis*. Two years after the *Flint Water Crisis* began, an Archive-It collection [33] for the story was created by Michigan State University (MSU). This potentially omits reports about the early stages of the water crisis. In fact, 81% of the MSU seeds were selected in 2016, two years after the state of emergency declaration of the water crisis. There are many reasons for the delayed creation of Web archive collections. One is the fact that many big stories such as the Flint story start small. Such stories are born into obscurity, and thus are not covered by major news outlets. This reduced exposure means archivists and curators may not know about the story until it creeps into the national spotlight once it becomes popular. Unfortunately, link rot persists from the time when the story starts to when it becomes popular. In fact, the Flint story was not covered by the national media until one year after the E. coli outbreak [31]. But local news media reported the Flint story from the beginning. The national media deserves criticism for the slow response in covering the Flint story, but any such criticism should be weighted by the fact that local and national news media have different priorities, and this is reflected by their respective news reports. We do not have the luxury of foreknowledge to determine what small stories will become big stories worthy of Web archive collections, and building a Web archive collection for every story is impractical. However, it is important to chronicle big stories from their small beginnings so as to provide preliminary context for their Web archive collections.

The absence of a Web archive collection could potentially impede the study of an event, especially one that occurred since the Web gained popularity, because we have ordained the Web as a primary historian and record keeper. Unfortunately, the Web forgets, and consequently, Web archive collections could serve as a means to reconstruct the historical record of important events. However, as we have seen, Web archive collections for important events are often absent (e.g., the *Stoneman Douglas Shooting* event). They are absent sometimes due to the lack of domain knowledge, and in other cases they may be created long after an event has occurred (e.g., the *Flint Water Crisis*), potentially omitting early events. These are just a few of the reasons we can attribute for the lack or lateness in the creation of Web archive collections; it is fair to expect that there more reasons. Irrespective of the reasons for why we do not have Web archive collections for many important events, the consequence is often the same - missing Web archive collections for important events. This

could adversely affect retrospective studies of events, by presenting an incomplete picture of an event, which may result in the establishment of a wrong conclusion. The shortage of curators can be reduced if we could automatically create seeds for Web archive collections. A natural question is: can we automate the seed generation process to bootstrap Web archive collections? In other words, can we develop and implement an algorithm for creating seeds for Web archive collections. The goal of this research effort is to address these questions.

1.3 AUTOMATING THE SEED GENERATION PROCESS

At the center of automatically generating seeds for stories and events is the issue of domain knowledge. Domain knowledge enabled the health experts at NLM to know what URLs are appropriate seeds for the *Ebola virus* Web archive collection. Domain knowledge qualifies an informed Arab resident in Tunisia or Egypt to be well-suited to create a Web archive collection for the *Arab spring* in Tunisia or Egypt, and is what qualifies a resident of the United States to create a Web archive collection for the 2008 or 2016 US Presidential Elections. Generating seeds for Web archive collections for arbitrary news stories and events, which is the intent of this research effort, requires arbitrary domain knowledge. We argue that it is currently impossible to automate this arbitrary domain knowledge since it requires being an expert in all things. However, arbitrary domain knowledge can be approximated by exploiting the collective domain expertise of Web users by using the collections they are already creating to generate seeds.

1.3.1 AUTOMATING THE SEED GENERATION PROCESS WITH SEARCH ENGINES

Web Search Engines (SEs) which are the primary means of discovery on the Web, prioritize recency, and thus, produce the most recent documents with respect to the time a query is issued [24]. For example, Table 1 (No. 1 – 5) shows a list of the first five URIs extracted from the Google Search Engine Result Page (SERP) for the query: “flint water crisis,” issued on November 7, 2018. As seen from items 1 and 2, Google returned two recent news stories² that were created the same day as the query issue date. Additionally, the search results included stories from 2016 – 2017, but the Flint water crisis began in 2014. Consequently, creating a collection from these search results may produce a collection

²<https://www.motherjones.com/politics/2018/11/dana-nessel-wins-michigan-ag-race/> and <https://www.beckershospitalreview.com/population-health/3-years-after-flint-newark-faces-water-crisis.html>

TABLE 1: A list of the first five URIs extracted from different collection building sources (Google and Twitter) and a Tweet reply thread [34], for the *Flint Water Crisis* story extracted 2018-11-07. The entries are sorted in reverse order of publication.

#	Pub. date	Page Title (URI)
Google		
1	2018-11-07	Michigan’s New Attorney General Wants to Shake Up the Flint Water Crisis Investigation (https://www.motherjones.com/politics/2018/11/dana-nessel-wins-michigan-ag-race/)
2	2018-11-07	3 years after Flint, Newark faces water crisis (https://www.beckershospitalreview.com/population-health/3-years-after-flint-newark-faces-water-crisis.html)
3	2017-03-28	Flint water crisis (https://www.nrdc.org/flint)
4	2016-04-20	Lead-Laced Water In Flint: A Step-By-Step Look At The Makings Of A Crisis (https://www.npr.org/sections/thetwo-way/2016/04/20/465545378/lead-laced-water-in-flint-a-step-by-step-look-at-the-makings-of-a-crisis)
5	2016-01-09	Flint water crisis (https://en.wikipedia.org/wiki/Flint_water_crisis)
Twitter Search (Top)		
6	2018-11-07	Michigan’s New Attorney General Wants to Shake Up the Flint Water Crisis Investigation (https://www.motherjones.com/politics/2018/11/dana-nessel-wins-michigan-ag-race/)
7	2018-11-07	The Flint Water Crisis May Finally Have A Champion In Dana Nessel (https://hillreporter.com/the-flint-water-crisis-finally-has-a-champion-in-dana-nessel-13381)
8	2018-11-01	EXCLUSIVE: Flint Water Declared 'Restored' After Michigan’s Environmental Agency Broke EPA Testing Regulations (https://medium.com/status-coup/exclusive-flint-water-declared-restored-after-michigan-s-environmental-agency-broke-epa-testing-3e2fc1f91a70)
9	2018-04-30	Support Gretchen (https://www.gretchenwhitmer.com/)
10	2017-03-16	Flint Water Crisis: What Happened and Why? (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353852/)
Tweet reply thread		
11	2016-01-13	Michigan governor orders national guard to assist in Flint’s water crisis (https://www.theguardian.com/us-news/2016/jan/13/michigan-governor-national-guard-flint-water-crisis-lead-rick-snyder)
12	2016-01-09	When money matters more than lives: The poisonous cost of austerity in Flint, Michigan (https://www.salon.com/2016/01/09/when_money_matters_more_than_lives_the_poisonous_cost_of_austerity_in_flint_michigan/)
13	2016-01-05	State of emergency declared over polluted drinking water in Michigan city (https://www.theguardian.com/us-news/2016/jan/05/lint-drinking-water-lead-pollution-michigan-governor-state-of-emergency)
14	2015-12-25	FOIA Request Shows Govt Lied About Lead in Water, Knowingly Poisoning Countless Children (https://www.alternet.org/news-amp-politics/foia-request-shows-govt-lied-about-lead-water-knowingly-poisoning-countless)
15	2015-12-15	Flint mayor declares ‘manmade disaster’ over lead-tainted water supply (https://www.theguardian.com/us-news/2015/dec/15/michigan-mayor-declares-manmade-disaster-lead-tainted-water-supply)

that disproportionately includes newer stories about the Flint water crisis, which may not report the early stages of the crisis. However, Google’s inclusion of the Wikipedia page for the *Flint Water Crisis*³ event increases the odds for the introduction of older content, but this is conditioned on whether a Wikipedia page for an event exists and is ranked within the top k pages a seed extractor visits.

Similar to Web SERPs, social media SERPs such as the Twitter SERP, are widely used to extract seeds URIs from tweets “as is.” This means the retrieved URIs are extracted from tweets returned by some Twitter filter (e.g., *top* tweets or *latest* tweets). For example, the URIs extracted from Twitter’s top vertical are produced by applying a filter based on Twitter’s notion of popularity (combination of top retweets, likes, freshness, etc.). For example, Table 1 (No. 6 – 10) shows a list of URIs extracted from tweets (top vertical) for the query: “flint water crisis.” Similar to Google (Table 1, No. 1), the first two URIs⁴ extracted from the tweets were created the same day as the query issue date (November 7, 2018). In fact, one of such stories titled *Michigan’s New Attorney General Wants to Shake Up the Flint Water Crisis Investigation* appears in Google and Twitter. The second story titled *The Flint Water Crisis May Finally Have A Champion In Dana Nessel*, reports on the same news event: the appointment of Dana Nessel as the new Michigan Attorney General. Additionally, Twitter’s top search included a marginally relevant URI (<https://www.gretchenwhitmer.com/>) because it was tweeted⁵ by a popular Twitter user (Hillary Clinton) in support of Gretchen Whitmer, the Governor of Michigan. The URI is only relevant when viewed in context of the tweet that embeds it. This highlights that we cannot blindly use URIs extracted from tweets as seeds. The application of filters is not necessarily a disadvantage, since it could reduce spam, however, Twitter’s filters do not have the editorial discretion expressed by human curators who manually select and collect URIs to generate seeds.

1.3.2 AUTOMATING THE SEED GENERATION PROCESS WITH MICRO-COLLECTIONS IN SOCIAL MEDIA

Users on social media sites such as Wikipedia, Twitter, Reddit, and Storify [36] routinely create and share narratives about news events consisting of hand-selected URLs of news

³https://en.wikipedia.org/wiki/Flint_water_crisis

⁴<https://www.motherjones.com/politics/2018/11/dana-nessel-wins-michigan-ag-race/> and <https://hillreporter.com/the-flint-water-crisis-finally-has-a-champion-in-dana-nessel-13381>

⁵<https://twitter.com/HillaryClinton/status/1056897327192043521>



Fig. 5: The Wikipedia page [35] for the MSD High School shooting, created the same day as the shooting event (February 14, 2018).

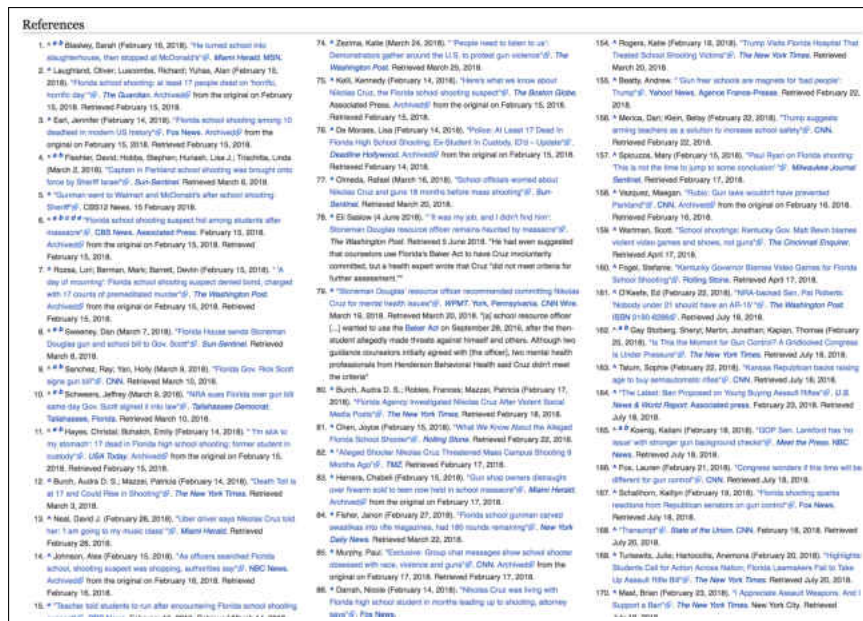


Fig. 6: References from Wikipedia MSD High School shooting page. As of July 2, 2020, it had 271 references. We propose extracting URIs from Micro-collections such as this to generate seeds.

stories, tweets, videos, etc. In this work, these narratives, whether shared on Wikipedia or Twitter, are called *Micro-collections*. It is important to note that seed generation systems that exclusively utilize Web and social media SERPs might be doing so without knowing what process generated the seeds, unlike Micro-collections which are created by social media users. We define Micro-collections as *social media posts that contain URLs that are gathered by humans as a demonstration of domain expertise and editorial activity, using the existing tools of social media platforms*. We consider Micro-collections as an important source for seeds because the effort taken to create Micro-collections is an indication of editorial activity, and thus presumably quality of the seeds. Web archive curators spend time selecting and filtering seed URI candidates. Similarly, social media users often perform similar tasks when faced with the decision of choosing what URIs to include in a “non-standard” social media post. For example, on the same day as the tragic *Stoneman Douglas Shooting* event, a Wikipedia⁶ page (Figure 5) was created for the event. Over two years after the event, the references (Figure 6) from the *MSD shooting* Wikipedia page had over 260 URLs pointing to news articles and other webpages related to the shooting event. Similarly, one day after the shooting event, a Twitter Moment [37] (Figure 7) was created. It consists of URLs of news stories as well as videos, images, and tweets about the event. As of July 2, 2020, there was no Archive-It collection about the event, thus URLs from the Twitter Moment (Figure 7) and Wikipedia references (Figure 6) may be used as seeds to bootstrap an Archive-It *Stoneman Douglas High School shooting* collection.

Figure 8a shows a story on Storify created January 2014 about the riots in Kiev, Ukraine. This was before the incident became a crisis in late February 2014 when Russia began the annexation of the Crimean Peninsula. In contrast, the Archive-It collection about the Ukraine conflict (Figure 8b) started in February 2014, and potentially omits some of the prelude contents in the Storify story (Figure 8a) which could be used to augment the Archive-It collection. Similarly, Table 1 (No. 11 – 15) shows a list of the first five URIs extracted from a Tweet reply thread [34] (Micro-collection) extracted with the query: “flint water crisis,” from the Twitter top search result page. This Micro-collection produced the oldest URIs⁷ compared to Google and Twitter. The stories titled *Flint mayor declares ‘manmade disaster’ over lead-tainted water supply*, and *FOIA Request Shows Govt Lied About Lead in Water, Knowingly Poisoning Countless Children*, highlight the early stages

⁶https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting

⁷<https://www.alternet.org/news-amp-politics/foia-request-shows-govt-lied-about-lead-water-knowingly-poisoning-countless> and <https://www.theguardian.com/us-news/2015/dec/15/michigan-mayor-declares-manmade-disaster-lead-tainted-water-supply>

17 people are dead after school shooting in Florida

US news · February 15, 2018

Authorities responded to reports of shots fired near Marjory Stoneman Douglas High School in Parkland, Florida. The local sheriff says there are multiple injuries and 17 people are dead. The shooter has been taken into custody.

22,498 Likes

Like Tweet



AP The Associated Press  @AP · Feb 14 2.9K 4.9K

BREAKING: Sheriff: Florida school shooter about 18 years old, not a current student, arrested without incident off campus.

 **David Ovalle**  @DavidOvalle305 · Feb 14 2.3K 4.3K

BREAKING: Florida school shooting suspect was ex-student who may have been flagged as campus threat. "We were told last year that he wasn't allowed on campus with a backpack on him."

Florida school shooting suspect was ex-student who was flagged as threat
A teacher at Marjory Stoneman Douglas High recalls a warning issued about ex-stud...
miamiherald.com

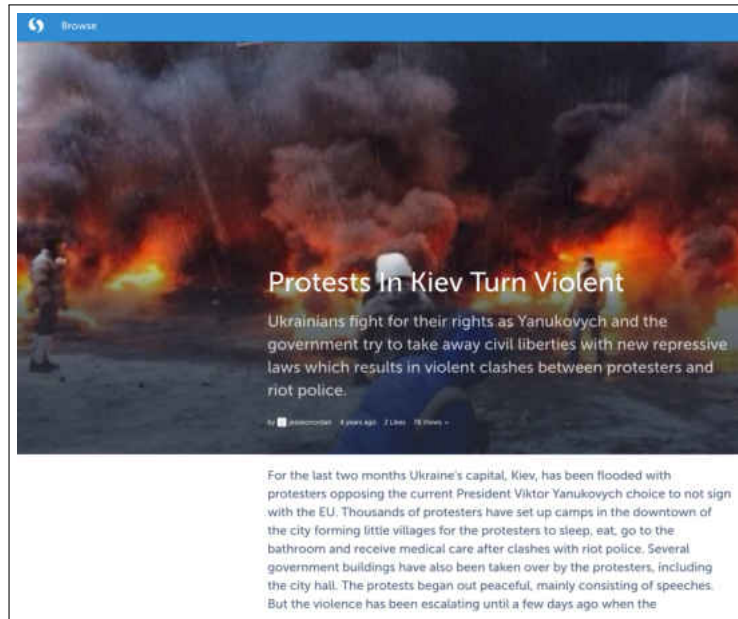
Fig. 7: A Twitter Moment [37] about the Stoneman Douglas High School shooting created the day after (February 15, 2018) the tragic incident. Social media Micro-collections such as this provides the opportunity for creating seeds to bootstrap archived collections. This is especially useful when no archived collection for the event exist; as of July 2, 2020, there was no Archive-It collection for the *Stoneman Douglas High School* shooting event. This screenshot has been edited to show more detail.

of the Flint story.

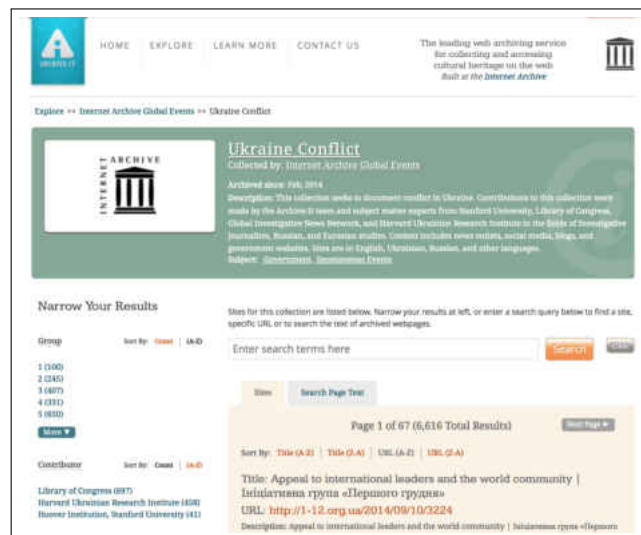
The *Ukrainian conflict* event highlights a common scenario in which users on social media express early interest and build Micro-collections for events before they gain prominence in the public discourse. This research effort will additionally explore finding these high-quality Micro-collections in social media to bootstrap archived collections. Kleinberg [38] introduced the concepts of *authorities* (information sources) and *hubs* (provide links to authorities) in the Web graph. Similarly, we consider Micro-collections as valuable *hubs* that could provide high-quality URLs that could be leveraged to generate seeds.

The users who create Micro-collections may not consider such activities as “seed generation,” or “collection building,” but it is fair to consider the creation of Micro-collections, a collection building process because of two reasons. First, generating seeds requires domain knowledge as we have seen from the NLM *Ebola virus* collection and the *Flint Water Crisis* collection. Similarly, the Micro-collections on social media such as Wikipedia and Twitter Moments created by various users are a demonstration of domain expertise. One is less likely to create a narrative (Micro-collection) about a subject by providing links to news articles and pictures to support an idea in the absence of domain knowledge. Therefore, the action of creating a Micro-collection is an expression of domain knowledge. Second, there are often a multitude of potential candidate URLs to include in Micro-collections. This means creators of Micro-collections rank the potential candidates and select a representative few. For example, there are potentially hundreds of thousands or millions of webpages that qualify as seeds for the *2016 US Presidential Election*. Even though a much smaller subset of these are visible to users based on the dynamics of the tools they use to discover content and their social media environment, users still have to sample from a larger pool of candidates to extract a smaller representative list of URLs to include in Micro-collections. This filtering activity often involves removing non-relevant or marginally relevant webpages, ranking the final candidates, and selecting a representative few. In other words, creating Micro-collections is an expression of editorial discretion. This editorial discretion is similarly applied in the seed generation process, where curators must decide what URLs should be seeds. For example, a curator might consider a seed of a high quality if it originates from an authoritative source and contains novel content.

Micro-collections created by social media users offer the opportunity for bootstrapping archived collections. Therefore, we propose a method of exploiting the collective domain expertise of Web users by using Micro-collections they are already creating to augment



(a) A story [39] from Storify: “Protests In Kiev Turn Violent,” published in January 2014. We propose extracting URIs from Micro-collections such as this to generate seeds.



(b) The *Ukraine Conflict* Archive-It collection [40] created February 2014.

Fig. 8: The Micro-collection from Storify (a) for the *Ukrainian crisis* event was created in January 2014 and highlights incidents such as riots before the event became a prominent news event. Russia began the annexation of Crimea in late February coinciding with the creation of the Archive-It collection (b). The Archive-It collection potentially omits some of the prelude contents in the Storify Micro-collection (a).

or bootstrap archived collections. In other words, the URLs extracted from such Micro-collections may serve as standalone seeds or augment curator-selected seeds for various news events. For example, Table 2 juxtaposes seeds from an Archive-It collection and URLs extracted from Reddit and Wikipedia⁸ for the *Ebola virus* topic. URLs from Reddit and Wikipedia can also be used to augment existing *Ebola virus* collections or bootstrap new ones. Since important events occur at a rapid pace, we cannot rely exclusively on archivists and curators for generating collections. Generating seeds from user Micro-collections on social media provides the opportunity for building a larger number of collections faster for important news events and for assisting archivists and curators in the collection building process.

1.4 RESEARCH QUESTIONS

The primary objective of this research effort is to automatically generate seeds for Web archive collections. Generating seeds requires domain knowledge, therefore, we propose addressing the domain knowledge problem (Chapter 1.2) by extracting URLs from Micro-collections generated by users on social media. This enables exploiting the collective domain expertise of social media users, thus removing the burden of encoding our automatic seed generation system with domain knowledge. The generated seeds may be crawled in the absence of curator-generated seeds to create Web archive collections for various stories and events, or may augment pre-existing curator-generated seeds. However, before generating Micro-collections, we must first identify them. This leads to our first research question:

- **RESEARCH QUESTION 1: How do we identify, extract, and profile Micro-collections in social media?**

Identifying Micro-collections makes it easier to extract them. Subsequently, it is important to establish profiles for the Micro-collections we find on social media, in order to facilitate describing them.

There are currently two popular methods for automatically or semi-automatically generating seeds. The first involves extracting seeds from SERPs. For example, to extract *Ebola virus* seeds from Google, one might issue the query: “ebola virus,” and collect the URLs from the top two pages. The second method involves extracting URLs from hashtags on Twitter. For example, to extract *Ebola virus* seeds from Twitter, one might extract URLs from the top 100 tweets surfaced with the hashtag #ebolavirus. In this research,

⁸https://en.wikipedia.org/wiki/Ebola_virus_disease

TABLE 2: Sample of seed URLs from Archive-It *Ebola virus* collection, URLs extracted from Reddit SERP (Search Engine Result Page) and comments for query “Ebola virus,” and URLs extracted from the references of the Wikipedia *Ebola virus* document.

Index	Title	URI
Archive-It (seed URLs)		
1	Eman Reports From Ebola Ground Zero...	blogs.plos.org/dnascience/2014/11/06/eman-reports-ebola-ground-zero/
2	Human rights and Ebola: the issue of quarantine...	blogs.plos.org/globalhealth/2014/11/ebola_and_human_rights/
3	2014-2016 Ebola Outbreak in West Africa...	www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html
4	#EbolaResponse (@ebola_response)...	twitter.com/ebola_response/
5	WHO — Situation assessments: Ebola virus...	www.who.int/mediacentre/news/ebola/en/
Reddit		
6	Liberia: Catholic Hospital Boss Tested Positive...	allafrica.com/stories/201407310957.html
7	Ebola plagues Africa nearly four decades...	america.aljazeera.com/articles/2014/8/1/ebola-explainer.html
8	Management of Accidental Exposure to Ebola Virus...	jid.oxfordjournals.org/content/204/suppl_3/S785.long
9	Analysis of patient data from laboratories during...	journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005804
10	Monkey Meat and the Ebola Outbreak in Liberia...	youtu.be/XasTcDsDfMg
Wikipedia		
11	Proposal for a revised taxonomy of the...	www.ncbi.nlm.nih.gov/pmc/articles/PMC3074192
12	Ebola outbreak in Western Africa 2014...	www.ncbi.nlm.nih.gov/pmc/articles/PMC4313106
13	Ebola data and statistics	apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest
14	WHO - Ebola outbreak 2014-2015	www.who.int/csr/disease/ebola/en/
15	Ebola virus entry requires the cholesterol...	www.nature.com/articles/nature10348

we propose a third method for extracting seeds - extracting seeds from Micro-collections. Therefore, it is pertinent that we compare the new method for generating seeds with the previous popular methods. Such comparison could enable us understand if the methods are similar, or the characteristics of the seeds generated with the different methods, and such information would be highly informative to future collection building processes. This leads to our second research question.

- **RESEARCH QUESTION 2: Do seeds from Micro-collections differ from seeds from SERPs and hashtags?**

Since we plan to generate seeds that can be used in addition to or in the absence of curator-generated seeds, it is important that the automatically generated seeds and the curator-generated seeds are comparable. Our functioning premise is that Archive-It seeds such as the *Ebola virus* and *HIV/AIDs* seeds (both created by NLM), and the *Flint Water Crisis* seeds (created by MSU) are gold standard seeds because they were created by experts. This may involve studying the structure of Archive-It seed collections [41]. This leads to our third research question:

- **RESEARCH QUESTION 3: How do we evaluate automatically-created collections with those generated by human experts in Archive-It?**

Addressing the third research question is critical since its solution can establish a method for evaluating our model for generating seeds automatically from Micro-collections. It however poses some challenges since it requires comparing collections that may cater to different needs, and there are many possible measures for comparing collections, so how does one narrow down this list to informative metrics that reflect if two collections are similar or dissimilar?

CHAPTER 2

BACKGROUND

In this chapter, we explore the prerequisite topics and concepts necessary to understand the remaining chapters. This begins with an introduction of the Web, and the means of discovering and preserving content on the Web through crawling strategies and Web archiving, respectively. Finally, we conclude by exploring some social media services that provide tools that encourage collection building activity.

2.1 THE WEB

The World Wide Web Consortium (W3C), the organization responsible for maintaining Web standards, defines the World Wide Web (WWW), or the Web as an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URIs) [42, 43]. The URL (Uniform Resource Locator), the most common form of URI, specifies the location of a resource, while the URN (Uniform Resource Name), a less common URI, is a location-independent identifier of a resource within a namespace. The URI is a superset of the URL and the URN. URIs are colloquially referred to as URLs, however for the remaining chapters, we identify resources with URIs and not URLs. Resources are usually in the form of webpages and are abstractions of entities (conceptual or physical) of informational value ranging from weather information to soccer game results. Since there are many possible resources on the Web, we need a means to identify them. Consequently, it is the primary task of URIs to identify resources. A single resource could manifest in different *representations* such as HTML or PDF documents. The act of retrieving a representation of a resource identified by its URI is called *dereferencing* the URI [44]. It is important to note that it is a representation of the resource and not the resource that is retrieved, even though it is common to say that the resource is dereferenced. Figure 9 illustrates the relationship between the URI, resource, and representation. To further explain the Web concepts, let us consider an example scenario [43] in which a person called Nadia wants to retrieve weather information before a trip to the Mexican City of Oaxaca.

To access a representation of the weather information resource of Oaxaca, Nadia performed the following operations:

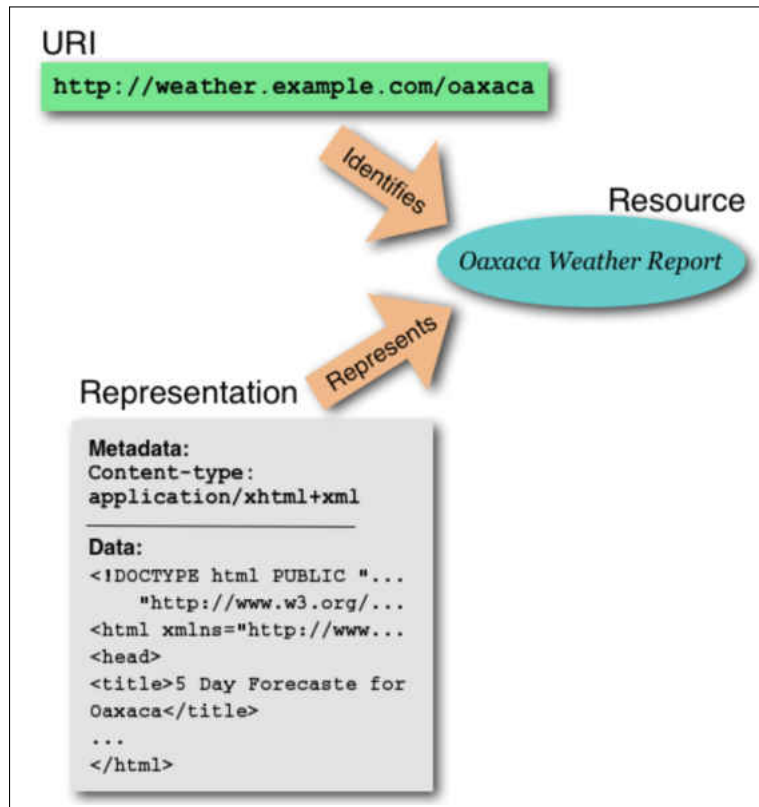


Fig. 9: W3C [43]: Illustration showing the relationship between URI, Resource, and Representation.

1. In order to interact with the Web, Nadia used a computer program called a *Web browser*. The Web browser acts on her behalf to retrieve information on the Web, and is thus called a *user-agent*.
2. Nadia typed the URI of the weather information resource: `http://weather.example.com/oaxaca`. Subsequently, the browser dereferenced the URI and received an HTML representation.
3. The browser takes the HTML representation of the resource and renders it showing a predicted weather report of sunshine at a temperature of 75°F. HTML (HyperText Mark-up Language) [45] is the most common representation of Web resources. It is a machine-readable code that Web browsers interpret and display as the webpages.

Nadia's browser is a *client*, and it requested the HTML representation of the Oaxaca weather report from a computer called a *server*, possibly housed thousands of miles away.

The client may request a different representation from the server, such as a PDF document, and initiate such a request through the process of *content negotiation* [46]. The Internet is the global system of interconnected computer networks that provided the channel of communication between Nadia’s computer and the Oaxaca weather report server. The Web is one of multiple applications that run on the Internet including email and file transfer. Similar to Nadia’s browser, there are many browsers that request weather reports from various servers over the Internet. The communication of the clients and servers over the Web is governed by a set of rules (*protocol*) called HyperText Transfer Protocol (HTTP) [46].

In addition to typing the URI of a webpage, e.g., `https://example.com/page.html`, one can visit the same webpage by simply clicking on a link embedded in a different webpage from a Web browser. In fact, browsing the Web often means following links from one website to another website. The link is the fundamental building block of the Web, which is composed of billions of webpages containing links to other webpages. The act of clicking a link translates to the act of dereferencing the URI into a representation. Browsing the Web enables the discovery of new resources and their respective webpages. Another prominent method of finding webpages does not involve clicking links, but instead typing a query into a search engine to surface webpages relevant to the query. Search engines such as Google provide links to documents relevant to a query, but this is only possible because all search engines possess an index generated by a *Web crawler*. The task of a Web crawler is to create an index that maps terms (e.g., “ebola”) to pages that include the terms (e.g., `https://www.cdc.gov/vhf/ebola/index.html`).

2.2 WEB CRAWLING

The basic Web crawling process (Figure 10) utilized by search engines to discover, save, and generate indexes for new URIs is outlined as follows. First, the Web crawler is provided with an initial list of URIs called *seeds*. Second, seeds are added to the crawl *frontier*. The frontier contains the list of URIs not yet visited. Third, the URIs in the frontier are dereferenced. Fourth, the resource representation (often HTML) is saved and processed as follows: URIs embedded in the HTML documents are extracted, and the downloaded documents are subsequently processed to generate the search engine index. Fifth, new URIs are added to the frontier, and old URIs (previously seen) are marked as visited, and used to update the index. The process restarts from the third step, and continues until a stop criteria is reached. The Web crawling process is a highly computationally and time intensive process,

and as a result, it is often performed in parallel. Web crawlers’ performance optimizations include reordering the crawl frontier, such that the most “important” pages are visited first [47, 48], and traversing the Web is done in a breadth-first manner [49, 50].

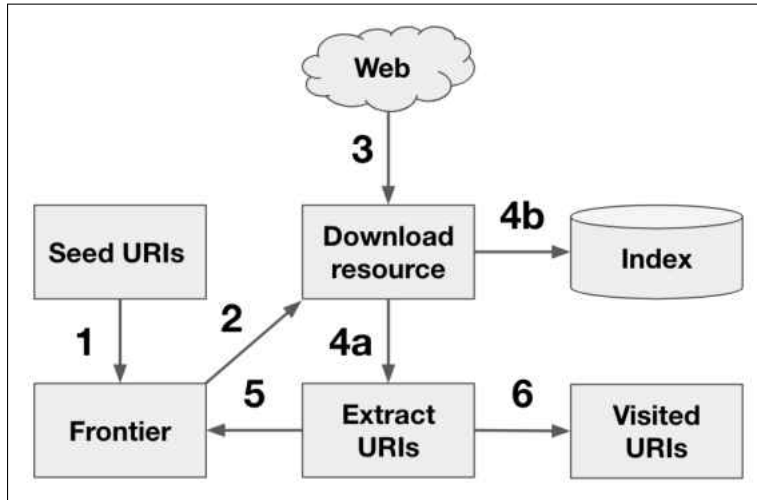


Fig. 10: Illustration showing the Web Crawling Process.

The first Web crawlers date back to the early 1990s [51, 52, 53] when the Web was small. In 1993, Matthew Gray at MIT wrote what is considered the first Web crawler - *Wanderer*, a Perl-based system that traversed the Web and indexed sites, and remained functional between June 1993 to January 1996 [54]. It was designed to discover new sites and measure the size of the Web. Shortly after *Wanderer*, *Jump Station* [51, 52] and *RBSE Spider* [55] were also released. Many of these crawlers such as *Jump Station* were restricted to indexing titles and headings of webpages due to limited resources. Similar to the previous early crawlers, 1994 saw the release of more Web crawlers such as Oliver McBryan’s *World Wide Web Worm* (WWW) [56]. The early Web crawlers used a set of seeds to collect various statistics about the Web and updated their respective indexes based on the information crawled.

In 1998, Brin and Page introduced Google [57], a large-scale Web crawler designed to address the scalability problem of a growing Web. At the center of Google is the *PageRank* algorithm which was designed to assign a quality rank for each webpage. PageRank calculates the probability of a user visiting a page based on the number of links that point to the page. This is based on the assumption that the higher the number of links to the webpage, the more important the webpage. Google’s approach significantly improved the quality of search results and set a new quality standard for search engines.

Unlike conventional Web crawlers designed to create indexes for search engines without taking the topics of webpages into consideration, focused crawlers (Figure 11) are designed to create collections of documents that are relevant to a predefined set of topics, e.g., *Aviation* or *Sports*. This is achieved by equipping a conventional Web crawler with a topic classifier. After a webpage is downloaded, the topic classifier decides if the webpage is relevant to the collection topic. If the webpage is relevant, the conventional crawling process continues, otherwise the webpage is discarded. Figure 12 illustrates the respective targets of conventional Web and focused crawlers.

The first focused crawler was introduced by Chakrabarti et al. [58, 59] in 1998. They defined the topic using a set of exemplar documents, and subsequently, the focused crawler was guided by a classifier that determined if an incoming document was relevant to the collection topic. Only documents relevant to the collection topic were included in the collection. Since the first focused crawler, there have been many variants of focused crawlers

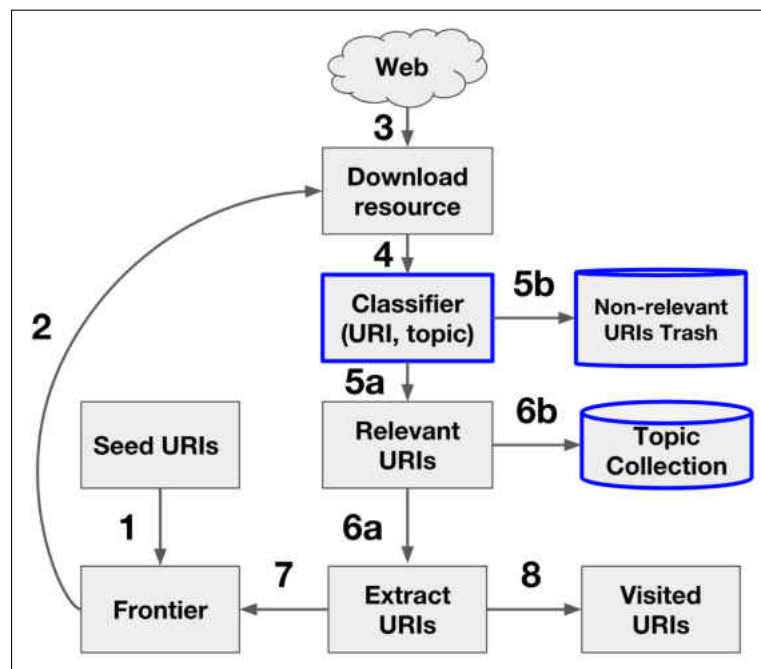


Fig. 11: Illustration showing the Focused Crawling Process. The blue annotation marks the new parts added to a Web crawler to convert it to a Focused Crawler.

[60, 61, 62, 63, 64, 65, 66, 67]. In general, focused crawlers keep their crawls focused by performing *link structure analysis*, *content analysis*, or a *hybrid approach* that combines the previous two methods [58, 68]. Link structure analysis is based on the idea that documents

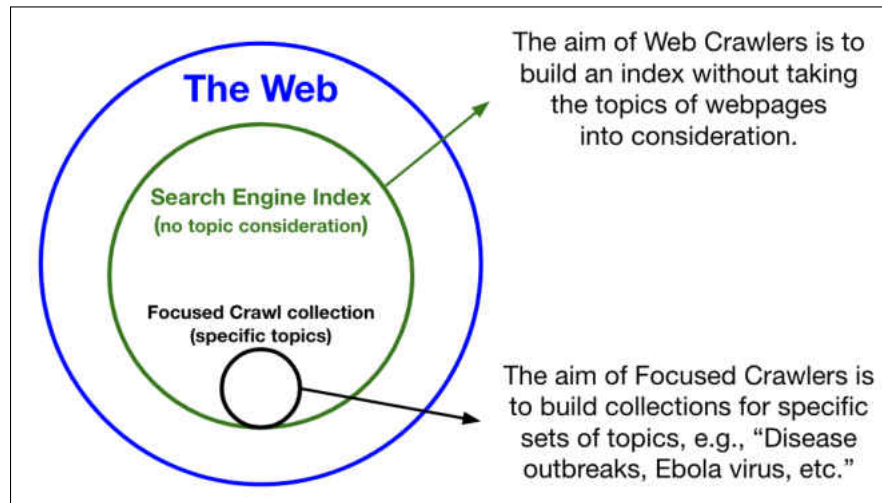


Fig. 12: Targets of Web and Focused Crawlers. Web crawlers used by search engines build (and update) indexes without taking the topics of documents into consideration, but focused crawlers focus on collecting documents that are similar to a narrow set of topics.

point to other documents similar to it [69]. This requires some notion of measuring similarity. Some methods employ a citation [70] or a co-authorship relationship [71] to quantify link similarity. Content analysis identifies similar documents based on the idea that similar documents share similar vocabularies [72].

Crawling and focused crawling are two primary processes of discovering new URIs on the Web. These two processes are highly computationally-intensive and often involve industrial-scale computer hardware. As a result, small-time operations that seek to crawl the Web are disadvantaged. In order to deal with the high cost of crawling, some have proposed crawling search engines.

2.3 CRAWLING SEARCH ENGINES

There are proposals to crawl search engines [73, 74, 75] as a means of augmenting existing collections. For example, the NASA Langley Research Atmospheric Data Center (ASDC) is a digital library that contains about two petabytes of earth science data. The ASDC has been utilized to produce scholarly work such as publications, webpages, visualizations, etc. However, the ASDC does not maintain information about these related scholarly products. Klein et al. [73, 74] proposed to remedy this by using the top search results from Google, Yahoo!, and MSN in order to augment the ASDC with related web resources. The search

results were retrieved with the use of the search APIs of Google, Yahoo!, and MSN.

The utility of crawling search engines is not limited to the augmentation of digital library collections. Crawling search engines also has utility in preservation. Klein and Nelson [76] presented a means of recovering missing web resources by proposing a method that partly relies on retrieving tags and link neighborhood lexical signatures of the missing resource from search engines. Similarly, McCown et al. [77] proposed *lazy preservation* as part of an effort to recover lost websites through the utilization of Web archives and search engine caches.

Crawling search engines results is highly useful, but not without limitations. Many research efforts that crawl search engines use search engine APIs to extract search results instead of scraping their Web User Interfaces (WUIs). This is partly because APIs are meant to be used by automated agents while WUIs are meant for humans. Unfortunately, studies [78, 79] have shown significant inconsistencies between API results and WUIs.

As we saw in Chapter 1, link rot (“resource not found”) is prevalent on the Web. This means crawling the Web to build an index, or focused crawling the Web to build a collection for a particular topic, or crawling search engines to augment existing digital library collections, are not sufficient. Consequently, conventional Web crawling is supplemented with preservation by Web archiving in order to mitigate link rot by preserving resources in Web archives.

2.4 WEB ARCHIVING

Web archiving is defined as the process of persistently collecting and preserving webpages in a Web archive. An archived copy of webpage may be viewed in place of a lost original copy, but this is only possible if the original webpage was saved. The Internet Archive founded by Brewster Kahle is currently the largest public Web archive. It has been collecting and saving public webpages since its inception (1996) and currently holds about 339 billion webpages. The Web archiving process is a complex one, and organizations must balance competing goals such as depth and width of coverage. Helen Hockx-Yu [80] provided a high-level summary of the Web archiving process as a composition of the following parts:

1. **Selection:** the decision-making process that determines what websites to archive.
2. **Harvesting (or crawling):** the automated process of downloading copies of websites, it also involves manual and automatic quality assurance.
3. **Storage:** the process of saving the downloaded websites on a storage medium that



Fig. 13: Helen Hockx-Yu [80]: Key Processes of Web Archiving.

ensures security and reliability. Archived websites are stored in one of two standard archival formats - ARChive (ARC) [81] or Web ARChive (WARC) [82] formats.

4. **Access:** involves providing access of archived websites to users.
5. **Preservation:** the Web standards, best-practices, and technologies needed to provide persistent access to the Web archives.

In addition to the Internet Archive, there are other public Web archives [83, 84], such as the UK Web Archive¹ and Icelandic Web archive². The standard way to view a snapshot of a webpage is to search for the URI of the website in a given Web archive. However, this approach is problematic because it requires one to have prior knowledge of the growing number of Web archives. The Memento Project³ proposed the *Memento* HTTP protocol [85] to remedy this problem by providing a means of linking the present (live) Web with the past (archived) Web, distributed among many Web archives.

2.4.1 THE MEMENTO FRAMEWORK

¹<https://www.webarchive.org.uk/>

²<https://vefsafn.is/index.php?page=english>

³<http://timetravel.mementoweb.org/>

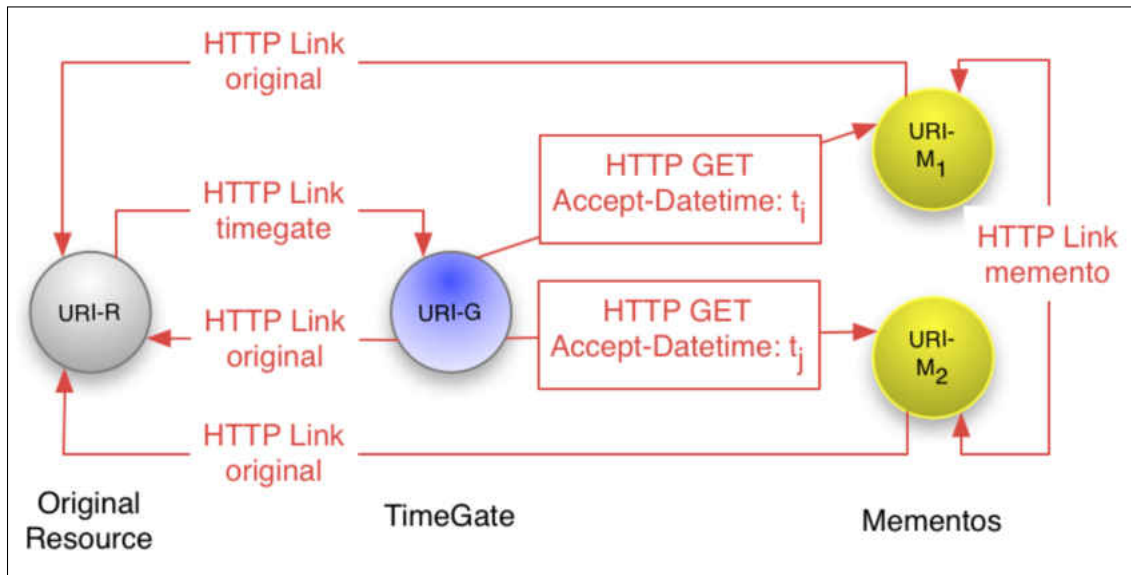


Fig. 14: Memento Framework [86]: Architectural overview of how the Memento framework allows accessing a prior version of a resource.

The Memento protocol (or Memento) (Figure 14) links the live and past Web by providing a client with content negotiation in the time dimension. As outlined in Figure 14, Memento links a URI-R (e.g., <http://www.cdc.gov/vhf/ebola/index.html>) on the live Web to its archived copy (e.g., <http://web.archive.org/web/20140325132254/http://www.cdc.gov/vhf/ebola/index.html>). Let us consider a brief outline of the primary components involved in the Memento framework:

1. **Original Resource (URI-R):** A resource as it exists or used to exist on the live Web is identified by the URI-R (e.g., <http://www.cdc.gov/>). This is the resource for which we seek a prior version.
2. **Memento (URI-M):** An archived copy of the original resource which is identified by its respective URI-M. Since a resource could have multiple copies, it could possess multiple URI-Ms. For example, the URI-M <http://wayback.archive-it.org/all/19961222063026/http://www.cdc.gov/> identifies an archived copy of the original resource (URI-R: <http://www.cdc.gov/>) from 1996.
3. **TimeGate (URI-G):** Given an input datetime, the TimeGate (e.g., <https://memgator.cs.odu.edu/timegate/http://www.cdc.gov/>) is a resource that uses content negotiation to provide the memento that is closest to the requested datetime.

4. **TimeMap (URI-T):** A TimeMap (e.g., <https://mimgator.cs.odu.edu/timemap/json/http://www.cdc.gov/>) is a machine-readable sorted list of URI-Ms. In other words, a TimeMap list the URIs (URI-Ms) of archived copies (mementos) of a given resource.

2.5 SOCIAL MEDIA TOOLS FOR COLLECTION BUILDING

The primary goal of social media tools such as Facebook and Twitter is to connect users. The operating theme for these social media tools might differ, such as connecting friends and families (e.g., Facebook) or connecting professionals in the same industry (e.g., LinkedIn), but fundamentally, social media tools provide the means of connecting users and provide a platform for the sharing of interests and ideas. Social media users actively create and share posts about important events. These posts often include hand-selected URIs (e.g., Figure 15) of news stories, images, videos, etc. Even though authors of the posts may not label their social media posts “collections” or “seeds,” these posts often involve two primary operations involved in traditional collection building: selection and filtering. Consequently, instead of relying exclusively on a few experts to generate seeds amidst a multitude of local and global stories and events, this work proposes leveraging the collective domain expertise of social media users by utilizing their social media posts to generate seeds. In this work, we refer to these posts as Micro-collections.

2.5.1 STORIFY

Storify⁴ was a social networking service launched in September 2010. Unfortunately, Storify went out of service on May 16, 2018 [89, 36]. It is however included in this section because it highly informed our research, since the posts (called *stories*) served as good examples of Micro-collections and showed collection-building activity in social media. Storify stories consist of hand-selected elements such as URIs of news articles, tweets, images, videos, etc., often embedded within a text narrative created by the author. AlNoamany et al. [90] conducted a study to understand the characteristics of Storify stories and discovered that popular stories were comprised of a median of 28 elements and a median of 12 multimedia resources (e.g., images and videos).

In Chapter 1, we saw an example of a story (Figure 8a) created when the protests in Kiev started. This story includes multiple URIs of news articles (e.g., [---

⁴<https://storify.com/>](http://</p>
</div>
<div data-bbox=)

Progressive Turnout Project @TurnoutPAC Follow

As we look ahead to 2019, we're counting down some of our favorites of 2018!

#10: Electing leaders like @AndyKimNJ and @SpanbergerVA07 that will fight for campaign finance reform!

Momentum builds for Dems to tak...
Momentum is building within the House Democratic Caucus to move aggressively on campaign finance
thehill.com

2:50 PM - 26 Dec 2018

4 Retweets 16 Likes

Tweet your reply

Progressive Turnout Project @TurnoutPAC · 27 Dec 2018

#9: We elected leaders who will prioritize affordable healthcare for all Americans.

Five health-care priorities for Dem...
Democrats won a majority in the House on Election Day, powered largely by their message on health ...
thehill.com

1 Retweet 8 Likes

Progressive Turnout Project @TurnoutPAC · 27 Dec 2018

#8: In 2018, we hired the largest Progressive Turnout Project staff to date! Our 400+ dedicated staff members were proud to help make the Blue Wave happen!

95 views 0:18 / 0:27

1 Retweet 3 Likes 11 Likes

Progressive Turnout Project @TurnoutPAC · 28 Dec 2018

#7: We elected Democrats who plan to restore voting rights and protect the right to vote!

Dems vow quick action to bolster ...

Join the discussion **BECOME A REDDITOR**

Posted by [u/LordVelaryon](#) 7 months ago

193 [Next Day Discussion] Post-Match Thread: Brazil 1-2 Belgium [2018 World Cup - Quarterfinals]

Post Match Thread

Brazil 1 - 2 Belgium

Quarter-finals

Stadium: [Kazan Arena](#) (Kazan, 42,873 Capacity)

Referee: [Milorad Mažić](#) (SRB)

FULL MATCH REPORTS:

- FIFA:** "It seemed written in the stars that Belgium would win. Every as Courtois showed just how good a goalkeeper he is, while the Red of luck in key moments. What's more, nothing seemed to go Neymar uncharacteristically uncertain in front of goal. The pain of Belgium's opponents in the Round of 16 at the 2002 World Cup enhances the j finals in Russia that little bit further."
- The Guardian:** "In years to come, when this stadium is a crumbling sit in almost empty stands, hear the wind whisper across the marsh believe what they hear is the ghosts of giants. In three games Kazan of 11 World Cups. First Germany went, insipid against South Korea. epic, Argentina were blown away by France. And then fell the biggest outwitted and outbattled by Belgium, who will face France in Tuesd
- BBC:** "Energy and belief is coursing through a Belgium team packed winners, and many are starting to seriously ask whether this squad finally about to deliver. If there was a touch of fortune about the first Ferdinandinho's arm and into the net, the second was beautifully craf four times at this tournament but his run which led to De Bruyne m behold, the striker receiving the ball inside his own half before turni. superb run. On a glorious night in Belgium football history, there w defender Thomas Meunier picked up another booking and will be su
- Kicker:** "The next top favourite has to drop out of the World Cup in out record world champions Brazil 2-1 in Kazan on Friday evening. 1 Americans had made life difficult for themselves - and late missed t controversial penalty-kick scene was also significant..."
- MARCA:** "A good reflection should also make Neymar, an extraordin end devoured by the character. He alternates brilliant moments witl controls with protests and unjustifiable attitudes. A mixture of god c for now does more harm than good. He leaves the World Cup being in the box than for his football and his goals. And it's a shame. And so good that he was about to score the goal of the draw in the 93'. I was blocked by Courtois with a spectacular flight. On the shore, Bra. Neymar finished."
- L'Equipe:** "After Messi and Cristiano Ronaldo's exits, Neymar proba World Cup and then winning his first Ballon D'Or. The World Cup, fo year. The Brazilian prodigy has also left the tournament. While he d, in the game, his overall performance was disappointing. The PSG st. fracture to his fifth metatarsal right foot in February, managed to r World Cup. Would a 100% Neymar have allowed Brazil to go all the haircut, then for his exaggerated falls, he certainly didn't experiece hoped for."
- La Gazzetta dello Sport:** "Belgium takes Brazil out and the World C Roberto Martinez's team will face France in the semifinals, on the ot Russia, England and Sweden remain. For the fourth consecutive Wor continent will raise the cup. At the end of another exciting match, cc Belgium beat Brazil 2-1, deadly in the first half and heroically resisti returned to the semifinals for the first time since 1986 and in this Wl goals, a record for them in the competition. Brazil falls unexpectedly and Messi and the fight for the Golden Ball remains very open."

POST-MATCH INTERVIEWS & QUOTES

[Full Official FIFA Post-Match Press Conference.](#)

Title:

(a) Subset of a series of tweets [87] by @TurnoutPAC about the 2018 US Midterm Elections.

(b) A Reddit post [88] by LordVelaryon about the 2018 FIFA World Cup. This post has been edited to show more details.

Fig. 15: A pair of social media posts consisting of multiple hand-selected URIs. The authors of posts such as these may not consider their posts as Micro-collections or the URIs as seeds, however, these posts exemplify collection building activity.

www.theguardian.com/world/2014/jan/22/ukraine-protests-three-dead, and <http://www.theguardian.com/world/2014/jan/22/ukraine-opposition-leaders-meet-president-protests-fatal>) and tweets (<https://twitter.com/kgorchinskaya/status/425996193936592896>, <https://twitter.com/carlbildt/status/428123839047147520>, and <https://twitter.com/SingingDDS/status/427978707886551040>) that chronicle the early stages of the crisis. We consider examples such as Figure 8a as Micro-collections and propose extracting URIs from them to generate seeds for Web archive collections, as a means to exploit the domain knowledge of social media users.

Even though Storify is out of service, the features it offers could potentially guide us to find alternative services [36]. Fortunately, there are other social media sites that offer similar services analogous to Storify’s stories. Let us consider the most prominent examples.

2.5.2 WIKIPEDIA

Wikipedia⁵ was launched in January 2001 as a free multi-lingual online encyclopedia edited by the public. The English Wikipedia currently has 5.7 million articles with an average of 560 articles added daily [91]. In addition to Wikipedia’s vast body of information about Arts and Science, it also includes articles about news events. For example, as we saw in Chapter 1, a Wikipedia page⁶ was created for the *Stoneman Douglas High School Shooting* event (Figure 5) the same day as the event. Similarly, the Wikipedia Ebola virus disease page⁷ was created on December 12, 2003, 10 years before the 2014 outbreak. Wikipedia editors often cite the sources of information used to create the articles in the article reference section. For example, both the Marjory Stoneman Douglas (MSD) High School shooting and Ebola virus Wikipedia page references contain over 230 URIs of news and scholarly articles related to the events, and thus offer the opportunity for generating seeds.

2.5.3 TWITTER

Twitter⁸ was launched in July 2006 as a social networking service in which users communicate primarily through short messages called *tweets*. A tweet may contain a combination of text, images, URIs, etc., but the service imposes a character limit on tweets. The original character limit was 140, but this was extended to 280 characters on November 7, 2017 [92].

⁵<https://www.wikipedia.org/>

⁶https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting

⁷https://en.wikipedia.org/wiki/Ebola_virus_disease

⁸<https://twitter.com/>

A single user on Twitter is connected to two categories of users: those the user *follows* and those *following* the user. Also, Twitter provides multiple means for users to engage with individual tweets. For example, a tweet may be *replied to*, *liked*, and *retweeted*. Twitter provides multiple tools that encourage collection building such as Twitter Moments, Twitter threads, and Twitter conversations.

Twitter Moments

Twitter Moments⁹ was launched in October 6, 2015 [93, 94] as a service that collects and shares tweets of noteworthy events as they unfold. A collection of tweets is called a *moment*. The staff at Twitter create the moments visible on the Twitter Moments website, however, ordinary users are also given the capability of creating moments. As we saw in Chapter 1, a Twitter moment¹⁰ (Figure 7) was created a day after the MSD shooting event, and it consists of URIs of news stories as well as videos, images, and tweets about the event. Twitter moments are simply a collection of tweets for a given topic, and since tweets may include URIs, the URIs may serve as seeds.

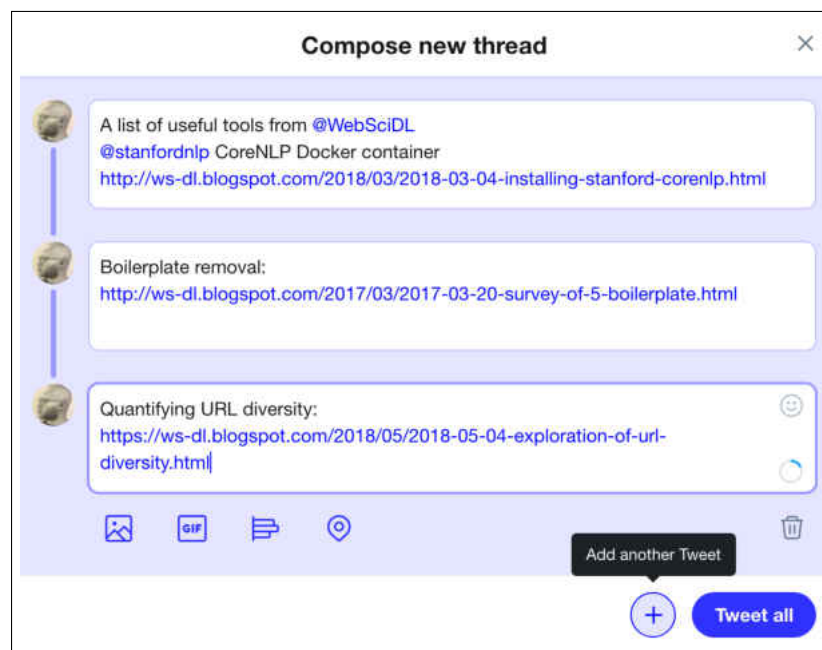


Fig. 16: *Nice Threads*, a feature for creating a collection of tweets by clicking the plus sign.

⁹<https://twitter.com/i/moments>

¹⁰<https://twitter.com/i/moments/963863619271254016>

Tammy Duckworth @SenDuckworth · 17 Feb 2017
Last night, I spoke for an hour on the Senate floor against Scott #Pruitt's nomination to lead the #EPA

99 134 1.1K

Tammy Duckworth @SenDuckworth · 17 Feb 2017
As someone who fought to defend this nation, I've seen firsthand the price we pay for our dangerous dependence on foreign oil

5 29 93

Tammy Duckworth @SenDuckworth · 17 Feb 2017
We should be encouraging the use of USmade renewable fuel&protecting #RFS not appointing people like #Pruitt who helped big oil sue the #EPA

6 26 78

Tammy Duckworth @SenDuckworth · 17 Feb 2017
As a mom of a toddler, I was shocked that #Pruitt was unaware that there is no safe level of lead for children at his confirmation hearing

13 60 129

Tammy Duckworth @SenDuckworth · 17 Feb 2017
Lead in our water supplies is a serious problem for kids in IL which is why we need an #EPA that will proactively prevent crises like #Flint


13 70 254

Tammy Duckworth @SenDuckworth Follow

Unfortunately, #Pruitt's record of filing lawsuits undermining the #EPA's ability to carry out its mission doesn't inspire much confidence

1:57 PM - 17 Feb 2017

Fig. 17: A tweet reply thread [95] about the *Flint Water Crisis* from Senator Tammy Duckworth consisting exclusively of text (no URIs or images). Reply threads are formed by replying to each preceding tweet, and thus provides an implicit means of creating a collection of tweets.


 **March for Science**  @ScienceMarchDC · 30 Mar 2017

Dr. Mona Hanna-Attisha is joining us as an honorary co-chair at the [#ScienceMarch](#) on 4/22! [#DefendTruth](#) [#ScienceNotSilence](#)



**“WE MARCH FOR SCIENCE
SO THAT SCIENTISTS HAVE
THE FREEDOM, LIKE I DID,
TO SPEAK OUT AND TO
CONTINUE TO MAKE THE
WORLD A BETTER PLACE.”**

Dr. Mona Hanna-Attisha
Pediatrician In Flint, MI
Honorary Co-Chair

MARCH FOR SCIENCE | APRIL 22, 2017



3 202 501

 **March for Science**  @ScienceMarchDC [Follow](#)

.@monaHannaA discovered that the untreated [#Flint](#) tap water created a lead poisoning crisis: [#NationalDoctorsDay](#)

 **Opinion | Will We Lose the Doctor Who Would Stop the Next...**
With walls and bans, our country will keep out talent, drive and creativity.
nytimes.com

10:43 AM - 30 Mar 2017

Fig. 18: A tweet reply thread [96] about the *Flint Water Crisis* from *March for Science* includes a single URI. The URIs embedded in reply threads may serve as seeds.



(a) The first subset of the tweet reply thread.



(b) The second subset of the tweet reply thread.

Fig. 19: A pair of three tweets that are part of a reply thread [34] from *Doing Things Differently* about the *Flint Water Crisis*. This reply thread spans over 2.5 years, and consists of 74 tweets (as of October 29, 2018) each containing a URI.

Twitter reply threads and conversation threads

The character limit of tweets restricts the amount of content that can be included in a tweet. This was especially the case when the 140 character limit was active. To overcome

this restriction, Twitter users often form a *reply thread* (officially known by Twitter as “Thread”) of tweets by replying to each preceding tweet. This causes the thread consisting of multiple individual tweets to be viewed at once instead of as a single tweet. Tweet reply threads provide an implicit method of building a collection of tweets, which may contain text content exclusively (Figure 17), a single URI (Figure 18), or multiple URIs (Figure 19). These hand-selected URIs embedded in tweet reply threads may serve as seeds.

In response to user behavior of creating reply threads, on December 12, 2017, Twitter announced a new feature (*Nice Threads* [97, 98]) that allowed users to create threads explicitly (without replying to tweets) at the push of a button (Figure 16). In this work, we call threads created using the explicit method introduced by Twitter, *conversation threads* to distinguish them from threads created implicitly (reply threads). However, there is no UI difference between explicitly and implicitly-created threads.


2.5.4 FACEBOOK

Facebook¹¹ was launched on February 4, 2004 as a social networking site that links communities of users called *friends*. When two users agree to be friends on Facebook, their messages, called *posts* become visible to one another, and they may interact by *liking*, *commenting*, *sharing*, etc. Posts from users have multiple privacy settings that control their visibility. For example, if a post is set as public, it is visible to the public. The visibility can also be restricted to a user’s friends. Unlike Twitter, Facebook does not impose a character limit on posts. Similar to other social networking sites, Facebook users often create posts about important news events. These posts often contain URIs of news articles, images, videos, etc., and thus make good candidates for seed extraction. For example, Figure 20 is one of such posts about the *#StopTheSoot* movement. The post [99] links to a 14-minute video [100] from Vice that exposes illegal refineries in the Niger-Delta region of Nigeria that contribute to the air pollution.

2.5.5 REDDIT

Reddit was launched on June 23, 2005 as a social networking site with an emphasis on web content rating. Reddit users post URIs for various topics and other users comment, vote up/down, or share the post. Reddit is organized into communities called *subreddits* based on

¹¹<https://www.facebook.com/>

 **Asemeyibo Buowari-Brown** is at Niger Delta. ⋮
 April 7 · Port Harcourt, Nigeria · 🌐

#StopTheSoot


If this is not sick then I don't know what else is. Commandant Haruna is definitely not fit for public office. If you watch the documentary and still feel the noise being made about the #StopTheSoot campaign is rubbish then I rest my case.




I therefore call on all well-meaning Nigerians, particularly the public figures, personalities, celebrities, social media influencers, news outlets to watch this documentary by @giannatoboni of HBO and lend their voices to this insane act.

This documentary shows the devastation that is ongoing with irresponsible handing of the illegal refineries situation in the Niger Delta. Two wrongs don't make a right and my heart bleeds...

Please spend 15mins to watch it for yourself and make up your mind.

Full length documentary 👉 https://youtu.be/vAgw_Zyxn0



   16

30 Comments 124 Shares 3.7K Views

Fig. 20: A Facebook public post from *Asemeyibo Buowari-Brown* about the *StopTheSoot* movement. The post links to a 14-minute video about illegal refineries operating in the Niger-Delta of Nigeria. Such refineries contribute to the air pollution in the region. This post has been edited to show more content.



Fig. 21: A Reddit post [101] from *jazir5* about Ebola virus vaccines. This post links to five authoritative sources that discuss the promise of immunity provided by Ebola virus vaccines.

topics. For example, the *news*¹² subreddit caters to news content and the *science* subreddit¹³ caters to science content. Similar to other social networking sites like Twitter and Facebook, Reddit users often create posts about important stories that contain URIs to news articles, images, videos, etc. For example, Figure 21 illustrates a comment from a Reddit user in response to a post of a URI <https://www.nature.com/articles/d41586-017-08664-w> titled: “Ebola survivors still immune to virus after 40 years,” from *Nature International Journal of Science*. The comment includes the following five URIs linking to articles about the promise of immunity provided by Ebola virus vaccines:

1. <http://www.who.int/en/news-room/detail/23-12-2016-final-trial-results-confirm-ebola-vaccine-provides-high-protection-against-disease>
2. <https://www.nytimes.com/2016/12/22/health/ebola-vaccine.html>
3. <https://www.nature.com/news/ebola-vaccine-approved-for-use-in-ongoing-outbreak-1.22024>

¹²<https://www.reddit.com/r/news/>

¹³<https://www.reddit.com/r/science/>

4. <https://www.niaid.nih.gov/diseases-conditions/ebola-vaccines>
5. <https://www.nih.gov/news-events/news-releases/experimental-ebola-vaccines-elicited-year-long-immune-response>

2.6 CHAPTER SUMMARY

We began this chapter by introducing the Web as an informational space of resources identified with URIs. The URIs may represent abstract or concrete entities and manifest in different representations, such as HTML, retrieved through a process called dereferencing. This was followed by an introduction of Web crawling as a means of discovering URIs that enables the building of search engine indexes. Focused crawling was similarly introduced as a means of discovering and saving URIs that are relevant to a specific set of topics. Next, we saw that partly due to the computational cost associated with crawling, there are proposals to crawl search engines. Next, we saw that Web archiving preserves Web resources by adding preservation to the Web crawling process. Finally, we concluded by showing that social media services such as Twitter provide services that encourage collection building, even though it is not addressed as such, and we showed that the Micro-collections from social media can provide us with a means to exploit the domain knowledge of social media users, through the generation of seeds from URIs in Micro-collections.

CHAPTER 3

RELATED WORK

This chapter explores other research work that informs ours, as well as the similarities and differences of our research with others that fall within the scope of collection building, seed generation, and collection evaluation.

3.1 COLLECTION BUILDING

Collection building refers to the generation of a set of documents relevant to a predefined topic such as the *#NODAPL* protests in the Standing Rock Indian Reservation, North Dakota, United States. Many research efforts exploit focused crawlers or variants of focused crawlers for collection building, but this is not always the case. Let us consider various research efforts that use, and subsequently, do not use, focused crawlers for collection building.

3.1.1 COLLECTION BUILDING WITH FOCUSED CRAWLERS

Bergmark [102] used the Mercator crawler [103] as a focused crawler for building collections by downloading webpages and subsequently classifying them into various topics in science, mathematics, engineering and technology. Her collection building approach is outlined as follows. First, a query is issued to a search engine, and a *centroid* is generated from the search results. The centroid, which is a list of representative seed URIs, is used to initialize the crawl frontier. Second, a crawl is issued, and only documents that exceed a given distance threshold are included in the collection. The distance metric used was a combination of cosine correlation, term vector, and vector space models.

Farag et al. [104] introduced the Event Focused Crawler, a focused crawler for events that uses an event model to represent documents and a similarity measure to quantify the degree of relevance between a candidate URI and a collection. An event is represented as a triple - *Topic*, *Location*, and *Date*. This is based on the definition of an event as something that happened in a certain place at a certain time (e.g., a shooting). The topic is represented as a vector created by extracting the top k keywords from a vocabulary extracted from the set of seed URIs. The location consists of a set of location entities (e.g., *New York City*)

frequently seen in the seed webpages. The date of the event is supplied by the user or extracted automatically from a set of seed webpages. Using the event model, consider the following event models for the *December 2015 San Bernardino Shooting* event,

Topic: shooting, shooter, ..., etc.
 Location: San Bernardino, California
 Date: 2015-02-12

and the *March 2016 Terrorist attack* in Brussels, Belgium

Topic: terror, attack, explosion, ..., etc.
 Location: Brussels, Belgium
 Date: 2016-03-22

Similar to Farag et al., Risse et al. [105, 106] introduced a new crawler architecture based on the ARCOMEM¹ project. Instead of the conventional crawling of all webpages, ARCOMEM strives to perform a semantic crawl of only webpages that capture *community memory*. Community memory involves webpages related to *events* and *entities* such as persons, locations, and organizations. This was achieved by extending the traditional crawl specifications that rely exclusively on seeds, to a hybrid *semantically-enhanced* specification that includes semantic information of the crawl intent such as entities and topics.

In the ARCOMEM project, Web and social media content are crawled independently, which could lead to a large delta (time gap) between when the collection and the embedded contents are fetched. Gossen et al. [107] proposed reducing the delta, and thus improving the freshness of the content. They proposed considering topical as well as temporal aspects when building collections in order to build collections with fresh content. This was achieved through the introduction of *iCrawl*, an integrated focused crawler that combines social media crawling and focused crawling of Web content to build topic-based collections. *iCrawl* utilizes social media content to guide the focused crawler toward fresh and relevant content. In this work, they estimated freshness F_P of a page P as the time difference between when a page is fetched t_f and estimated creation date of the page t_c ($F_P \approx t_f - t_c$). Similarly, they defined the relevance of a page and used the relevance score of pages and their corresponding outlinks to prioritize the visitation of pages, or determine what pages should be visited. To determine the relevance score of a given document, first, the crawl specification was represented as a *reference vector*. The crawl specification consists of an initial list of seeds, social media

¹ARCOMEM From Collect-All ARchives to COmmunity MEMories, <http://www.arcomem.eu/>

TABLE 3: Gossen et al. [108]: Exemplary scopes used in a sub-collection specification. This list is not exhaustive.

Scope	Type	Description
URL	List of URLs	Documents that need to be in the sub-collection
Domain	List of domains	Domains that the sub-collection should be restricted to
Time	Time interval	Relevant timeframe
Keywords	List of keywords	Descriptive keywords for the sub-collection topic
Event/Entity	List of knowledge base references	Entries in a knowledge base such as FreeBase [109] that are the topic of the sub-collection
Size	Number of documents	Target size of the sub-collection

queries, and keywords specified by a user. Second, the document to be evaluated was represented also as a vector called the *document vector*. Third, the cosine similarity was computed between the reference and document vectors. Once the relevance of pages in the crawl queue went below a predefined threshold, the crawl was stopped.

3.1.2 COLLECTION BUILDING WITHOUT FOCUSED CRAWLERS

Focused crawlers are often used (but not always) to generate collections. Gossen et al. [108] proposed a methodology for extracting Web archive collections focused on specific topics and events (called a *topic and event focused sub-collection*). A *topic and event focused sub-collection* is defined as a collection of documents in a Web archive collected using a *sub-collection specification* (Table 3). The *sub-collection specification* is a list of scopes that define how a sub-collection is generated. A sub-collection is generated as follows:

1. Choose a base Web archive W
2. Create a sub-collection specification CS based on your need
3. Select an algorithm A that supports the scope specified in the CS
4. Run A over W using CS as an argument

TABLE 4: Gossen et al. [110]: Examples of temporal event characteristics.

Event	Type	Duration	Lead time	Cool-down time
Olympic games	Recurring	2 weeks	Weeks	Days
Federal election	Recurring	1 day	Months	Weeks
Fukushima accident	Non-recurring	1 Week	-	Months
Snowden leaks	Non-recurring	1 day	-	Years

5. The result of the previous step is the sub-collection C

The method described above for extracting a sub-collections was proposed as an iterative process, which allows for the modification of the sub-collection specification (CS') or sub-collection generation algorithm (A'). The modification of the specification leads to a new sub-collection C' . Our research differs from Gossen et al. in two major ways. First, Gossen proposes generating collections from within the Web archives, but we propose generating seeds from the live Web to create Web archive collections. Second, Gossen proposed running an algorithm A over a sub-collection specification CS on a Web archive W to generate a sub-collection C . This is analogous to a clustering techniques that create buckets of items that are similar based on a predefined similarity criteria. This means the decision of whether a URI belongs in a sub-collection is encoded in the specification of an algorithm. However, in this work, we leverage the judgment of humans on social media. In a similar work, Gossen et al. [110] adapted some portions of the *topic and event focused sub-collection* in a method to extract event-centric documents from Web archives based on a specialized focused extraction algorithm. This was applied to a German Web archive covering a 19-year period for the extraction of event-centric collections for events such as the *Iraq war*, *Costa Concordia grounding*, and the *German federal elections (2002 - 2013)*. This work characterized an event as something that happened at a certain date (e.g., an accident) or time interval (e.g., a sports tournament). They defined two broad kinds of events based on time: *planned* and *unexpected*. Just as the name implies, planned events are events expected to occur at a particular time, e.g., elections. For these kinds of events, especially those that are recurring (e.g., FIFA World Cup tournaments), relevant documents often appear in advance of the actual begin time of the event during the event *lead time*, and documents are often continuously published after the event completion during the event *cool-down time*. However, for unexpected events, especially those that are non-recurring such as a terrorist

attack, documents related to the event are published from the start time of the event and the cool-down time of the event. There is no lead time for these events. Table 4 summarizes the temporal characteristics of events. The goal for the event-centric extraction process is, given an event input and a Web archive, generate an interlinked collection of documents relevant to the input event that meet the *collection specification*. The collection specification consists of the topical and temporal scopes and is defined as follows:

- **Topical scope:** One or more topical reference documents, and zero or more representative keywords.
- **Temporal scope:** Time span of the event (including the start and end dates) $T_e = [t_e^s, t_e^e]$, and time duration of the lead time (T_l) and the cool-down time (T_r).

Algorithm 1 Gossen et al. [108]: Event-centric Collection Extraction

Input: Collection Specification CS , $targetSize$

Output: Document collection c , excluded URLs $missing$

```

 $q \leftarrow \text{priorityQueue}(\text{seedUrls}(CS)); c \leftarrow \{\}; missing \leftarrow \{\}$ 
while not isEmpty( $q$ ) and  $|c| < targetSize$  do
   $url \leftarrow \text{pop}(q)$ 
   $v \leftarrow \text{resolveSnapshots}(url, CS)$  {Find all snapshots of  $url$  in  $c$ }
  if  $v = \emptyset$  then
     $missing \leftarrow missing \cup \{url\}$ 
  else
     $v_i \leftarrow \text{selectSnapshot}(CS, v)$ 
     $c \leftarrow c \cup \{v_i\}$ 
     $out \leftarrow \text{extractOutlinks}(v_i) - \text{seenUrls}$  { $\text{seenUrls} = c \cup missing$ }
     $\text{insert}(q, out, \text{relevance}(v_i))$  {Insert outlinks into queue according to relevance}
  end if
end while

```

Following the definition of the collection specification, Algorithm 1 describes the process of generating an event-centric collection. The differences of our research with Gossen’s previous work [108] (methodology for extracting *topic and event focused sub-collection*) transfer to this work. Additionally, this work requires the builder of the collection to possess domain knowledge about the lead and cool-down times, a requirement we do not impose on the user of our system, but one that can be included if such information is available. Regardless of the

differences between our approaches for generating collection, the previous two works from Gossen et al. inform our work, especially the dynamics of time in generating a collection.

Most focused crawling is performed on the live Web. Unfortunately, the live Web is plagued by link rot and content drift, consequently, Klein et al. [25], similar to Gossen et al. [110], demonstrated that focused crawling on the archived Web results in more relevant collections than focused crawling on the live Web, for events that occurred in the distant past. Additionally, similar to this work, Klein et al. proposed extracting seeds from external references contained in the Wikipedia page of an event. However, instead of utilizing the live version of the Wikipedia page, they proposed using the version of the Wikipedia page that corresponds with the datetime after which the edit frequency drastically decreases.

Similar to Gossen et al., Nanni et al. [111, 112] presented an approach for extracting event-centric sub-collection from Web archives. Their method extracts documents not only related to the input event, but also documents describing related events (e.g., premises and consequences). For a given event, they identified relevant concepts and entities from a knowledge base, and detected the mentions of the entities in documents. More specifically, given an event v embodied by a Wikipedia page w (e.g., *2014 Orange Revolution Wikipedia page*: https://en.wikipedia.org/wiki/Orange_Revolution), and a corpus C (e.g., New York Times Corpus), their method produces a collection of documents in C (event-collection) that are relevant to the event. Their method can be summarized in seven steps (Figure 22):

1. **Initial document retrieval:** An initial set of documents D is generated by populating the set with all documents in C that mention the event name (e.g., “Orange revolution”).
2. **Entity candidate collection:** A set of potentially relevant entities E is generated as follows. First, entities extracted from D (initial set of documents) using TagMe2 [113] are added to E . Second, entities from the outlinks of the Wikipedia page w (page corresponding to the event) are added to E in the entity query feature expansion process [114, 115].
3. **Entity ranking:** Entities in E are ranked as follows. First, every entity $e_i \in E$ that corresponds to a Wikipedia page (and DBpedia entity) is represented by a vector representation (knowledgebase embedding) calculated from its RDF graph [116]. Also the event is similarly represented by its knowledgebase embedding. Second, entity-event relatedness is calculated by computing the cosine similarity between the entity and event embeddings. Third, the entities are ranked based on their cosine similarity

scores.

4. **Entity-context passage collection:** For all entities $e_i \in E$, the passage $p(e_i)$ with the highest relatedness to the entity e_i is extracted from the Wikipedia page w of the event. The relatedness between an entity and an event was calculated by computing the cosine similarity of the entity and event GloVe [117] word embeddings (semantic vector representations).
5. **Embedding representation:** The set of entities E and set of all contextual passages P are projected into an embedding space to generate their latent feature vectors, GE and GP , respectively. This is achieved by computing the element-wise averages of the embeddings of E and P .
6. **Entity-query feature expansion:** The initial event query q is expanded with multiple vector expansion models such as the *Place*, *Ent-TFIDF*, *EvAsp-TFIDF* expansions, etc. The *Place* expansion involves the expansion with only the location entity. The *Ent-TFIDF* expansion involves expansion with the top 10 related entities from E . The *EvAsp-TFIDF* expansion involves expansion with the words from the contextual passages P of the top 10 related entities.
7. **Query processing and supervised document ranking:** The query q is normalized as follows. First, entities (e.g., “Yulia Tymoshenko”) are tokenized into word components (e.g., “Yulia” and “Tymoshenko”). Second, the words are represented as vectors. Third, documents in C are ranked based on their cosine similarity scores (query vs. document vector).

Nanni et al.’s method of generating event-centric collection for events with a Wikipedia page is similar to ours, however, the exclusive use of Wikipedia for defining an event may be restrictive for the following reasons. The existence of a Wikipedia page favors older events and popular events. Even though we may choose to build collections for popular events that have Wikipedia pages, this is not always the case. Some events start small and may not have a Wikipedia page, especially events that happen outside the countries with the most Wikipedia editors (such as United States, Germany, and Russia [118]). This means the proposed method cannot be applied to new events without a Wikipedia page. Consequently, we propose to consider other social media sources for generating collections.

3.2 SEED SELECTION

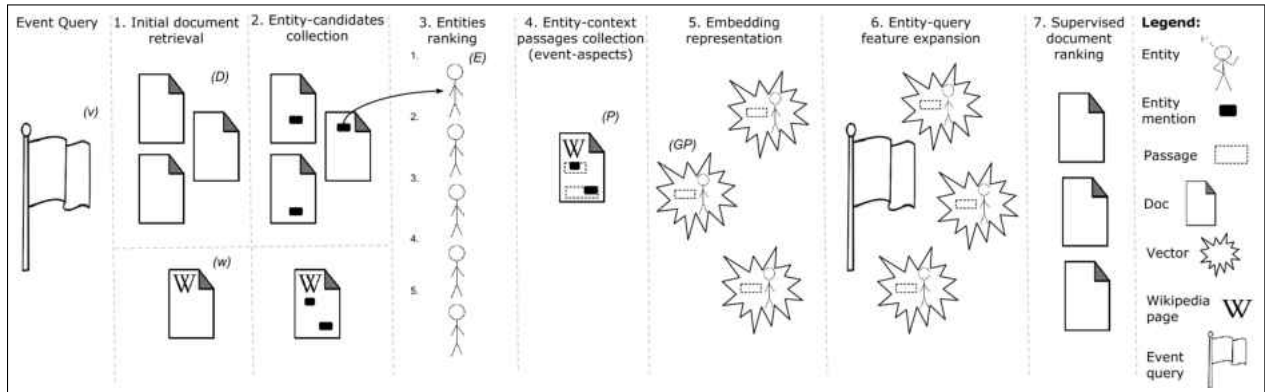


Fig. 22: Nanni et al. [112]: Overview of the method to extract event-centric sub-collections from Web archives

Selecting good seeds is challenging and has not been extensively studied. Collection building researchers often acknowledge the importance of selecting good seeds, and admit its link to the performance of their systems, but often they pay more attention to the mechanisms of building the collection and not seed selection. Collection building efforts mostly utilize search engines and social media (e.g., Twitter and Wikipedia) as sources of seeds (Table 5). The challenge of selecting good seeds is embodied in the idea that it is difficult to define “good.” This challenge is captured by Bergmark’s statement [102]: “It is unclear what makes a good seed URL, but intuitively it should be rich in links, yet not too broad in scope.” Zheng et al. [119] argued that the seed selection problem for Web crawlers is not trivial, and proposed different seed selection strategies based on PageRank, number of outlinks, and website importance. They also showed that different seeds may result in collections that are considered “good” or “bad.” While there have been efforts made to automatically generate seeds, many of these methods (e.g., Prasath and Öztürk [120]) target generating seeds for Web crawlers that build indexes for search engines, and not seeds for focused crawlers or Web archive collections.

Du et al. [121] proposed a customized method of generating seeds for focused crawlers based on the personal information of a user. They proposed creating a user-interest ontology that captures the interests of a user based on past Web usage. Subsequently, the user-interest ontology is used to expand the user query in order to extract additional seeds from a search engine. Since this method depends on historical use information, its performance is tied to the availability of such historical data, which might be lacking due to the absence of domain knowledge.

Social media, such as Twitter, is a popular source for extracting seeds. As part of the Crisis, Tragedy, and Recovery Network project², Yang et al. [122] proposed using URIs found in tweet collections as seeds to bootstrap Web archiving tasks quickly for sudden emergencies and disasters. Their prototype system for building Web archives with minimum human input by extracting seeds from tweets is summarized as follows. The input to the system comes from a tweet archive database. Next, a URL extractor periodically extracts URLs from tweets. The URLs extracted from tweets are stored in a URL database. Next, a Heritrix crawler is issued to crawl each URL in the URL database, and the crawled data is stored in the WARC format. Priyatam et al. [123] proposed extracting diverse seeds from tweets in a Twitter URI graph for the Web crawlers of digital libraries such as CiteSeerX [124] which offers specialized search for computer science articles. The process begins by extracting tweets that contain URIs with domain-specific queries. Second, an undirected unweighted graph is constructed such that the nodes represent the URIs and an edge is constructed between a pair of nodes if they are similar beyond a threshold. Similarity was calculated using four approaches: the *Content*, *URI*, *User*, and *Zero* approaches. In the *Content* approach, two URIs were considered to be similar if tweet texts that contained both URIs overlapped (Jaccard index) beyond a threshold. The *URI* approach measured similarity based on the level of overlap (Jaccard index) of the 4-grams of the URIs. In the *User* approach, two users were considered similar if at least one of them had retweeted the other’s tweet. *Zero* similarity meant no URIs were considered similar; this was the baseline similarity approach. Third, the graph generated in step two is passed into a diversification engine that returns k diverse URLs. The authors conclude that similarity calculated using a combination of *Content*, *URI*, and *User* approaches produced the best results. Even though this work does not target the generation of seeds for collections of stories and events, which is a focus of our work, the notion of diversity of seeds is adopted in our work (Chapter 6.1).

Table 5 is a summary of the already discussed collection building efforts from multiple perspectives including build target, focused crawling, seed creation, seeds source, and output. The table contrasts previous collection building research with ours by showing that our method automates seed creation.

Search engines and social media (e.g., Twitter) are two dominant sources of seeds for collection building efforts that use focused crawlers. In this work, we do not consider using focused crawlers to crawl seeds to discover more relevant URIs, but focus on how seeds can be generated by exploiting social media Micro-collections. This is different from the

²CTRNet: <http://www.ctrnet.net/>

TABLE 5: Summary of **Previous Research** and **This Research**'s approach toward collection building.

#	Research	Build Target	Focused Crawling	Seed Creation	Seeds Source	Output
Previous Research						
1	Bergmark [102]	Live Web	Yes	Semi-automatic	Search Engine	Collection
2	Farag et al. [104]	Live Web	Yes	Manual	N/A	Event Collection
3	Risse et al. [105, 106]	Live Web & Social Media	Yes	Manual	N/A	Collection
4	Gossen et al. [107]	Live Web & Social Media	Yes	Manual	Social Media	Collection
5	Gossen et al. [108]	Web Archive	No	Manual	N/A	Collection
6	Gossen et al. [110]	Web Archive	No	Manual	Wikipedia References	Event Collection
7	Klein et al. [25]	Web Archive	Yes	Semi-automatic	Wikipedia References	Event Collection
8	Nanni et al. [111, 112]	Web Archive	No	Semi-automatic	Wikipedia Document	Event Collection
This Research						
1	Nwala et al.	Live Web	No	Automatic	Social Media	Collection

Legend

Key	Properties
Build Target	Live Web/Web Archive/Social Media
Focused Crawling	Yes or No: was collection built with a focused crawler
Seed Creation	Manual/Semi-automatic/Automatic
Seed Source	Service from where seeds were extracted
Output	Topic Collection: collection of documents related to topic, and Event Collection: collection of documents related to an event

conventional use of search engines and social media for generating seeds in the following ways.

First, we do not use search engines (e.g., Table 5, No. 1) such as Google to generate seeds because the search results are highly sensitive to when a query is issued; search engines tend to provide the most recent URIs corresponding to the time a query is issued. This phenomenon is explained further in Chapter 4.1.

Second, efforts that utilize Twitter (e.g., Table 5, No. 4) to generate seeds (primarily with hashtags) tend to extract URIs from tweets “as is.” This means the retrieved URIs are extracted from tweets returned by some Twitter filter (e.g., *top* tweets or *latest* tweets). For example, the URIs extracted from Twitter’s top vertical are produced by applying a filter based on Twitter’s notion of popularity (combination of top retweets, likes, freshness, etc.).

Our proposed method of extracting seeds from Micro-collections relies on what users have collected, and not what a service decides to provide by the application of a filter. However, in this work, we compare the various strategies for generating seeds (Chapter 5.2).

Wikipedia references (e.g., Table 5, No. 6 & 7) are another popular source for generating seeds, and we consider these Micro-collections. Unfortunately, Wikipedia pages do not exist for many important events. For example, on August 20, 2013, a gunman entered an Atlanta elementary school with an assault-style rifle with the intent to kill. Fortunately, a school clerk (Antoinette Tuff) talked to him and convinced him to stay with her. The gunman was eventually arrested without causing any physical harm to anyone. There is no Wikipedia page for this incident [125, 126] probably because there were no casualties. In this work we do not rely exclusively on one source for seeds (from Micro-collections) such as Wikipedia references, but propose extracting seeds from Micro-collections from other social media such as Facebook and Reddit.

3.3 COLLECTION EVALUATION

We want to evaluate the collections we build. This need arises because there are different strategies for collecting seeds to generate collections. Some strategies involve extracting seeds from search engines such as Google [102, 24, 127], and others involve extracting seeds from social media such as Twitter [122, 123]. It is crucial to understand the properties of seeds extracted from various media; this is especially important since systems such as Google and Twitter are black boxes. Therefore, it is important to compare seeds generated using different strategies, as this could inform the decisions made in order to generate

collections. For example, due to the fast decay rate of finding news stories on the Google SERP (Chapter 4.1), one might consider Wikipedia to get early seeds for a long-running popular event. Collection evaluation offers methods for quantitatively characterizing the collections we build. This quantification provides a means to not only characterize individual collections but can also serve as a way to compare collections.

Collection evaluation dates back beyond the Web, to book collections in libraries. In 1974, Bonn [128] presented different quantitative methods for evaluating various library collections and expressed the need for library collections to be varied in order to fulfill the needs of various academic programs. Some of the methods discussed by Bonn for evaluating library collections include:

1. *Compiling various statistics on the library holdings, use, and expenditure:* This includes measuring the total volumes of the reference books in the library, volumes added per year, subject balance, unfilled requests, interlibrary loan request, circulation, expenditure, etc.
2. *Checking list, catalogs, and bibliographies:* This method of evaluating libraries involves checking the quantity of the library's holdings that is present in a list, such as the *Books for College Libraries* [129]. Additionally, the catalogs of important libraries such Harvard's *Lamont* and Princeton's *Julian Street* libraries are used.
3. *Conducting surveys of the users of the library:* This involves evaluating a library by taking into consideration the opinions of the users of the library such as faculty and researchers, students, and the general public.

In the 1980s, the Research Libraries Group (RLG), a consortium of libraries in the U.S., published the RLG six (0 – 5) collecting levels [130, 131] to quantify the strength of collections. In summary, level 0 means the library collection is out of scope with respect to a subject, and level 5 means the collection is comprehensive. Table 6 summarizes the RLG collecting levels. To further clarify the RGL six, language suffixes (Table 7) were included. For example, a library is assigned “3F” if it collects at the instructional support level for Spanish of Venezuela. More recently (2004), Lesniaski [132] provided a simplification of *White's brief tests* [133] (comparing a short list of items to a library's collection) in order to make the test more adaptable by smaller college libraries. Additionally, he expressed the idea that there is not a single meaning of a “good” library collection since the meaning is defined by the user or target audience of the collection.

TABLE 6: The RLG Collecting Levels

Level	Definition
0	Out of scope
1	Minimal
2	Basic information
3	Instructional support
4	Research
5	Comprehensive

TABLE 7: The RLG Collecting Levels Language Suffixes

Suffix	Definition
E	For primarily English-language material
F	For selected foreign-language material
W	For wide selection of foreign-language material
Y	For material primarily in one foreign language

The questions proposed by the library sciences such as “How does one evaluate collection strength?” and “What is a good collection?” are applicable to the Web domain. Many of the solutions offered by libraries for quantifying collection strength can be summarized into two broad categories: *collection-centered* and *use-centered* [134]. Collection-centered methods include comparing a collection against an expert-provided gold standard bibliographical set. Use-centered methods include assigning the strength score to a collection based on circulation and interlibrary loan statistics, and patron surveys [135]. At Web scale, a gold standard is very often absent, so checking a collection against a predefined list is not practical, however, some solutions offered by libraries to these questions (quantifying the strength of a collection) could inform the Web domain through transformations [136].

There are a few studies that address collection evaluation in the Web domain that inform this research. Risse et al. [137] surveyed social scientists, historians, and legal experts in order to extract the requirements they find desirable for building collections. Some of the needs include topical dimension, time dimension, and the need to crawl social media sites. Topical dimension refers to the need to chronicle the evolution of an event over time.

The time dimension is related to the topical dimension, but addresses the need to capture documents as events unfold. Some real world events have well-defined times e.g., a sports event and elections. Archivists often need the crawl duration to encompass the real world event time frame. Social media is increasingly where the first reports of many events such as protests and popular uprisings unfold, consequently, it is important to include social media sources in some social media driven collections.

3.4 CHAPTER SUMMARY

In this chapter, we considered research efforts in collection building, seed generation, and collection evaluation that are similar, different, and inform this work. We began by exploring various collection building research that utilize focused crawlers to generate collections targeting the live Web or the archived Web. We showed that search engines such as Google, and social media platforms such as Twitter and Wikipedia are popular services for generating seeds, but established the distinction between existing methods of generating seeds from these sources and our Micro-collection approach. Next, we explored research efforts that do not utilize focused crawlers for collection building. Finally, we concluded by showing that collection characterization and evaluation has not been extensively studied in the Web domain unlike the library sciences domain, and explored the various methods for evaluating collections in both domains.

CHAPTER 4

SCRAPING SEEDS FROM SERPS

As we saw in Chapter 3, Search Engine Result Pages (SERPs) such as the Google SERP are known for their high quality results, and thus are often used to generate seeds. Collection building often begins with a simple Google search to discover seeds. This can be done by issuing queries to Google and extracting URIs from the SERP (Figure 23). For example, the following are two possible candidate URIs extracted from the Google SERP to include in a collection (or seed list) about the *Hurricane Harvey* (August 2017) event:

`http://www.cnn.com/specials/us/hurricane-harvey`

`https://www.nasa.gov/feature/goddard/2017/harvey-atlantic-ocean`

The URIs extracted from Google can serve as seeds that can be crawled to build collections in Archive-It, such as the Archive-It *2017 Hurricane Harvey* collection¹. SERPs provide an opportunity to generate seeds for news stories and events, and SERP results influence the nature of seeds generated from them.

Even though search engines generate high quality seeds, seeds generated from search results are highly sensitive to when a query is issued; search engines have a recency bias and tend to provide the most recent URIs corresponding to the time a query is issued. This is not a negative feature, but a design constraint influenced by the fact that users typically want the most recent results for some query. Search engines are made to meet the needs of such users, not support the needs of scholars with an interest in increasing recall. Some search engines such as Google provide a means to alter this default behavior through date range filters, however, it becomes harder to find older documents as time progresses because they must compete with newer documents for a fixed number of slots, and search engines prioritize recency. Although this phenomenon is known, in this chapter we provide the result of our study [24] to quantify it. Queries used to extract news stories are examples of *informational queries* [138], and we expect their SERP results to change as the news event evolves. It is important to quantify the magnitude of this change because in order to build a representative collection about an event, we ought to capture not just a slice of time,

¹<https://archive-it.org/collections/9323>

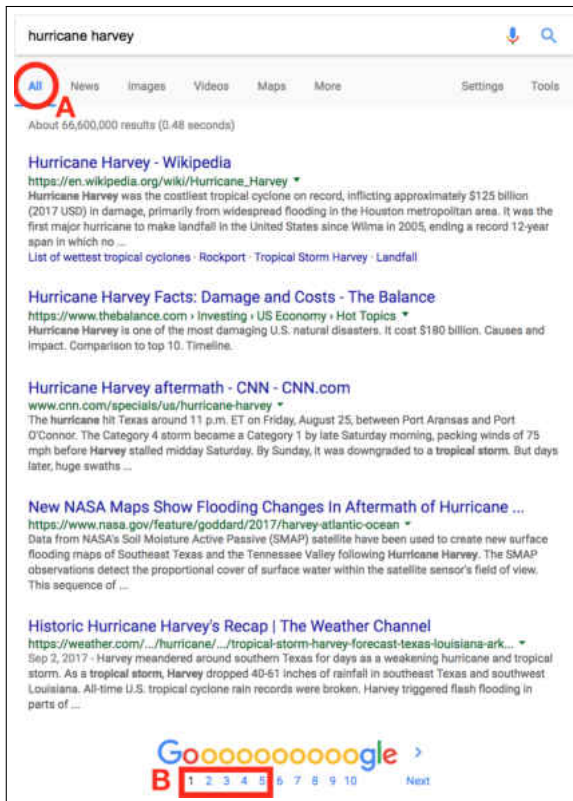
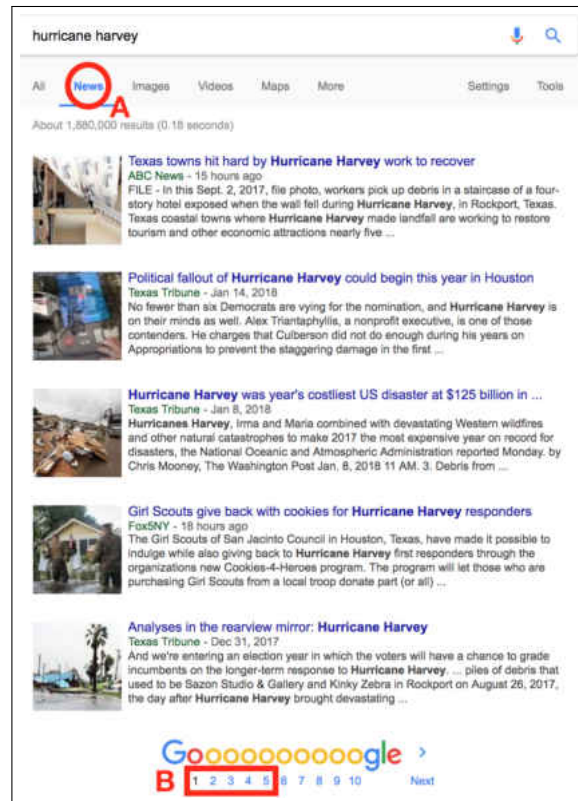
(a) The Google All (referred to as *General*) SERP.(b) The Google *News* vertical SERP.

Fig. 23: Google *General* (a) and *News* vertical (b) SERPs for the query “hurricane harvey.” Some links have been removed to enable showing more detail. For our experiment, links were extracted from the first five pages (annotation B) of both SERPs for each query.

but the various stages of the events [137] - oldest to newest. This discourages the exclusive use of SERPs for generating seeds for representative collections because SERPs favor recent documents. We expect the SERP results for *transactional* (e.g., “samsung galaxy s3”) or *navigational* (e.g., “youtube”) queries to be less transient [139], but such queries were not our focus.

To help understand and quantify the flux of search results, we conducted a study to assess how to “refind” the URIs of news stories on the Google SERP.

4.1 EXPERIMENT: REFINDING NEWS STORIES ON SERPS

Event-based collections often start with a Web search, but the search results you find on Day 1 may not be the same as those you find on Day 7. We studied seed collections

generated by extracting URIs from SERPs, specifically Google, in order to provide insight about the retrievability of URIs of news stories. SERPs are useful artifacts in their own right, and can be used for multiple activities such as classifying queries [140] as “scholarly” or “non-scholarly,” but this study focused on tracking the URIs of news stories on SERPs to answer the following questions.

- Can one “refind” the same URI of a news story (for the same query) from Google after a given time?
- What is the probability of finding a story on Google over a given period of time?

To address these questions, we issued seven queries to Google every day for over seven months (2017-05-25 to 2018-01-12) and collected links from the first five SERPs (Figure 23 annotation B) to generate seven collections for each query. The queries represent public interest stories that happened (or are happening) in different timelines.

- “healthcare bill”
- “manchester bombing”
- “london terrorism”
- “trump russia”
- “travel ban”
- “hurricane harvey”
- “hurricane irma”

We tracked each URI (extracted by issuing the respective queries) in all collections over time to estimate the discoverability of URIs from the first five SERPs of Google. Our findings (Chapter 4.2) suggest that it becomes more difficult to find the URI of a news story with the same query after a week, and almost impossible after a month.

4.1.1 DATASET GENERATION, REPRESENTATION, AND PROCESSING

The dataset extraction duration varied for the queries as outlined by Table 8. The dataset extraction process lasted from 2017-05-25 to 2018-01-12. For each query, we extracted approximately 50 links within `<h3>` HTML tags from the first five pages of the Google SERP from the default (*All*) and *News* vertical SERPs (Figure 23, annotation A & B).

To avoid confusion, in this chapter we refer to the *All* SERP as *General* SERP. The first five pages were considered in order to gain better insight about the rate of new stories across pages, as considering a few pages (e.g., 1 or 2) may present an incomplete view. In total, 73,968 (13,708 unique) URIs were collected for the *General* SERP and 77,634 (19,724 unique) for the *News* vertical SERP (Table 8). In previous work with the Local

TABLE 8: The *SERP-Refind* dataset [141] generated by extracting URIs from SERPs (*General* and *News* vertical) for seven queries between 2017-05-25 and 2018-01-12.

Collection (Query/ Topic)	Start date (duration in days)	News story count	
		General SERP count (unique count)	News vertical SERP count (unique count)
healthcare bill	2017-05-25 (232)	12,809 (2,559)	13,716 (3,450)
manchester bombing	2017-05-25 (232)	12,451 (1,018)	13,751 (1,799)
london terrorism	2017-06-04 (222)	10,698 (1,098)	10,450 (2,821)
trump russia	2017-06-06 (220)	12,311 (4,638)	13,728 (3,482)
travel ban	2017-06-07 (219)	12,830 (2,849)	13,439 (2,815)
hurricane harvey	2017-08-30 (135)	6,666 (685)	6,450 (2,530)
hurricane irma	2017-09-07 (127)	6,203 (861)	6,100 (2,827)
Subtotal		73,968 (13,708)	77,634 (19,724)
Collections Total		151,602 (33,432)	

Memory Project (LMP) [142], we introduced a local news collection generator [143]. The local news collection generator utilizes Google in order to build collections of stories from local newspapers and TV stations for US and non-US news sources. Unlike LMP, in this work we did not restrict the sources sampled to local news organization, but still utilized

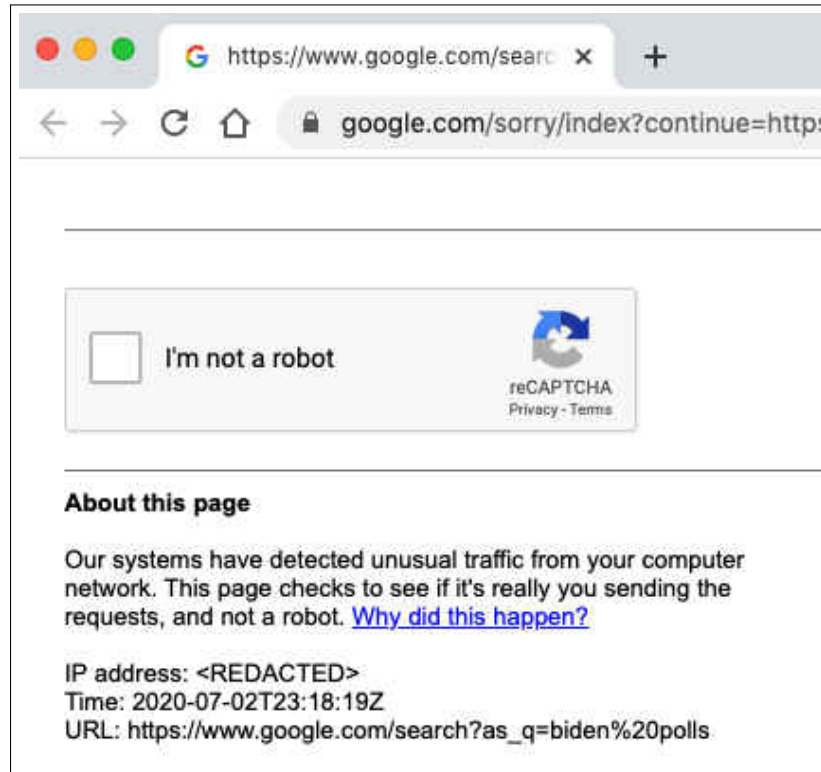


Fig. 24: A screenshot of the Google CAPTCHA page for query “biden polls,” triggered by searching 18 times (paginations counted), each time paginating to a maximum of page 20.

Google in order to discover seeds. The local news collection generator was used to scrape links from the Google SERP, and it was adapted to include the ability to extract all kinds of news stories from Google (not just from local news organizations). The Google search interface is meant for humans and not for robots, and it presents a CAPTCHA (Figure 24) when it is used too frequently in order to discourage automated searches. Consequently, the dataset collections were all generated semi-automatically with the use of the local news collection generator. The input provided to the extension was the query and the maximum number of pages to explore (five), and the output was a collection of URIs extracted from the SERPs. The URIs collected daily from the SERPs were represented as JSON files. For a single query, two JSON files per day were generated, each file represented the URIs extracted from the *General* SERP and *News* vertical SERP. This means for a given day, a total of 14 (two per query) JSON files were generated. Each URI in a JSON file included metadata extracted from the SERP such as the *page number* and the *rank* which is the position across all SERP pages. Additionally, each file included the date the data was generated.

At the center of the analysis was the ability to track the URI of a news story over time. URIs often have aliases (multiple URIs identifying the same resource). For example, the following pair (a and b) of URIs identify the same resource:

- (a) <https://www.redcross.org/donate/disaster-donations?campname=irma&campmedium=aspot>
- (b) <https://www.redcross.org/donate/disaster-donations>

As a result, we transformed all URIs before matching by trimming the scheme and all parameters from the URIs, using a method suggested by Brunelle et al. [144]. The parameters in URIs often express a reference source such as `origin` and `callback`, or session parameters such as `session`. The transformed version of the URI was used to track the individual news stories. Subsequently, for each news story we recorded all the dates and pages it was observed on the SERP.

4.1.2 MEASURES FOR TRACKING URIS ON SERPS OVER TIME

The following measures were extracted from the *SERP-Refind* dataset (Table 8) and provided insight on the discoverability of the URIs of news stories on the Google SERP.

Story replacement rate, new story rate, and page level new story rate

Given that at time point t_0 we observed a set of URIs for news stories u_0 and at time point t_1 we observed a set of URIs for news stories u_1 , then the story replacement rate at t_1 is given by Equation 1.

$$\text{Story replacement rate} = \frac{|u_0 - u_1|}{|u_0|} \quad (1)$$

For example, if we observed URIs $\{a, b, c\}$ at t_0 and URIs $\{a, b, x, y\}$ at t_1 , then the story replacement rate at t_1 is

$$\frac{|\{a, b, c\} - \{a, b, x, y\}|}{|\{a, b, c\}|} = \frac{|c|}{|\{a, b, c\}|} = \frac{1}{3} = 0.3.$$

We can see that at t_1 , one out of the three original URIs was replaced. Similarly, the rate of new stories going from t_0 to t_1 is given by Equation 2.

$$\text{New story rate} = \frac{|u_1 - u_0|}{|u_1|} \quad (2)$$

For example, if we observed URIs $\{a, b, c\}$ at t_0 and URIs $\{a, b, c, d, e\}$ at t_1 , then the new story rate from t_0 to t_1 is

$$\frac{|\{a,b,c,d,e\}-\{a,b,c\}|}{|\{a,b,c,d,e\}|} = \frac{|\{d,e\}|}{|\{a,b,c,d,e\}|} = \frac{2}{5} = 0.4.$$

At t_1 we observed new stories d and e . We calculated the story replacement rate and new story rate using different temporal intervals (daily, weekly, and monthly) individually for each of the first five pages of the *General* and *News* vertical SERPs. The daily story replacement rate indicates the proportion of stories replaced on a daily basis. This is similar to the daily new story rate because the SERP returns a similar number of results ($mean = median = mode = 10$ links, and $\sigma = 0.43$). The daily new story rate approximately indicates the rate of new stories that replaced previously seen stories on the SERP on a daily basis. The higher the story replacement and new story rates, the lower the likelihood of refinding previously seen stories.

Probability of finding a story

Given a collection of URIs for news stories for a topic (e.g., “hurricane harvey”), consider the URI for a story s_0 that was observed for the first time on page 4 of the SERP on day d_0 . We represent this as $s_0^{d_0} = 4$. If we find s_0 on page 2 on the next day d_1 and then it disappears for the next two days, we represent the timeline observation of s_0 as $\{4, 2, 0, 0\}$. Therefore, given a collection (e.g., “hurricane harvey”) of N URIs for news stories, the probability $P(s^{d_k})$ that the URI of a story s is seen after k days (d_k) is calculated using Equation 3.

$$P(s^{d_k}) = \frac{\sum_{n=1}^N T(s_i^{d_k})}{N}; T(s_i^{d_i}) = \begin{cases} 0 & ; \text{if } s_i^{d_i} = 0 \\ 1 & ; \text{if } s_i^{d_i} > 0 \end{cases} \quad (3)$$

The probability $P(s^{d_k} = m)$ that the URI of a story s is seen after k days (d_k) on page m , is calculated using Equation 4.

$$P(s^{d_k} = m) = \frac{\sum_{n=1}^N T(s_i^{d_k})}{N}; T(s_i^{d_i}) = \begin{cases} 0 & ; \text{if } s_i^{d_i} \neq m \\ 1 & ; \text{if } s_i^{d_i} = m \end{cases} \quad (4)$$

Distribution of stories over time across pages

For each story URI, we recorded the dates it was observed on the SERP. For each date, we recorded the page where the story was found. The collection of stories and the date/page observations were expressed using the notation introduced in Chapter 4.1.2. For example,

the following list of three URIs for news stories s_0 , s_1 , and s_2 were observed for the first time (first day - d_0) on pages, 4, 1, and 1, respectively. On the last day (d_3), the first story (s_0) was not seen on any of the pages ($s_0^{d_3} = 0$), however both the second (s_1) and third (s_2) stories were found on the first page ($s_1^{d_3} = 1$ and $s_2^{d_3} = 1$):

$$s_0 = \{4, 2, 0, 0\},$$

$$s_1 = \{1, 2, 0, 1\}, \text{ and}$$

$$s_2 = \{1, 1, 1, 1\}.$$

Overlap rate and recall

Given two sets of collections of URIs, A and B , the overlap rate $O(A, B)$ quantifies the amount of URIs common within both sets without considering the size disparities of the sets. This was calculated using the Overlap coefficient as follows: $O(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$. The standard information retrieval recall metric $r(A, B)$ for two sets of collections A and B with respect to A , quantifies the amount of stories present in A and B (as a fraction of A) was calculated as $r(A, B) = \frac{|A \cap B|}{|A|}$.

Our dataset was generated without setting any parameters on the Google SERP. However, the Google SERP provides a date range parameter that restricts the documents returned on the SERP to documents published within the date range. For example, setting the date range to 2017-06-01 and 2017-06-30, attempts to restrict the documents in the SERP to those published between June 1, 2017 and June 30, 2017. To assess the effect of setting the date range parameter on discovering older stories that fall within a specific timeframe, we took the following steps. First, from our original dataset, we selected five collections of stories for queries about topics that occurred before June 2017: “healthcare bill,” “trump russia,” “travel ban,” “manchester bombing,” and “london terrorism.” This set of five collections was called *June-2017*. Second, we removed all stories from *June-2017* that were not published in June 2017. Third, in January 2018, we issued the selected five queries to the Google SERP without setting the date range to generate five additional collections (from the first five pages). This set of five collection was called *Jan-2018* (control test collection). Fourth, we issued the same five queries to the Google SERP, but this time, we set the date range to 2017-06-01 and 2017-06-30, and extracted five collections. This set of five collections was called *Jan-2018-Restricted-to-June*. Finally, we calculated the overlap rate and recall between the *June-2017* and *Jan-2018*, as well as *June-2017* and *Jan-2018-Restricted-to-June* collections for the pairs of collections with the same query.

TABLE 9: Average **story replacement rate** for *General* and *News* vertical SERP collections. Column markers: **minimum⁻** and **maximum⁺**.

Collection	General SERP			News vertical SERP		
	Daily	Weekly	Monthly	Daily	Weekly	Monthly
healthcare bill	0.42	0.60	0.76	0.44	0.71	0.87
manchester bombing	0.27	0.39⁻	0.59⁻	0.31⁻	0.54⁻	0.76⁻
london terrorism	0.34	0.41	0.60	0.43	0.66	0.84
trump russia	0.54⁺	0.79⁺	0.92⁺	0.42	0.71	0.90
travel ban	0.43	0.63	0.82	0.45	0.62	0.83
hurricane harvey	0.21⁻	0.41	0.67	0.49	0.77	0.91
hurricane irma	0.27	0.44	0.73	0.57⁺	0.82⁺	0.92⁺

4.2 RESULTS

Here we present the results for each of the previously introduced measures.

4.2.1 STORY REPLACEMENT RATE, NEW STORY RATE, AND PAGE LEVEL NEW STORY RATE

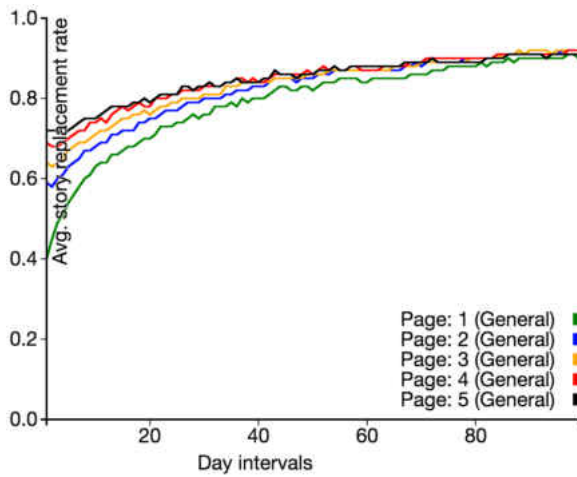
Tables 9 and 10 show the average story replacement rate and new story rate, respectively over time (daily, weekly, and monthly) for the *General* and *News* vertical SERPs. For both *General* and *News* vertical SERPs, we can see that the average story replacement rate was similar to the new story rate, and both increased with time. They also show that the story replacement and new story rates are strongly dependent on the topic. For example, the *Hurricane Harvey* natural disaster showed a lower daily average story replacement rate (0.21) and new story rate (0.21) compared to the *Trump-Russia* event, which maintained the highest daily (0.54), weekly (0.79), and monthly (0.92) average story replacement and new story rates (0.54 - daily, 0.78 - weekly, and 0.83 - monthly). Unlike natural disasters which have a well-defined timeframe, this on-going political event does not have a well-defined timeframe and as of January 2018, has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Similar to the *General* SERP, the average story replacement rate and new story rate for the *News* vertical SERP increased with time

TABLE 10: Average **new story rate** for *General* and *News* vertical SERP collections. Column markers: **minimum⁻** and **maximum⁺**.

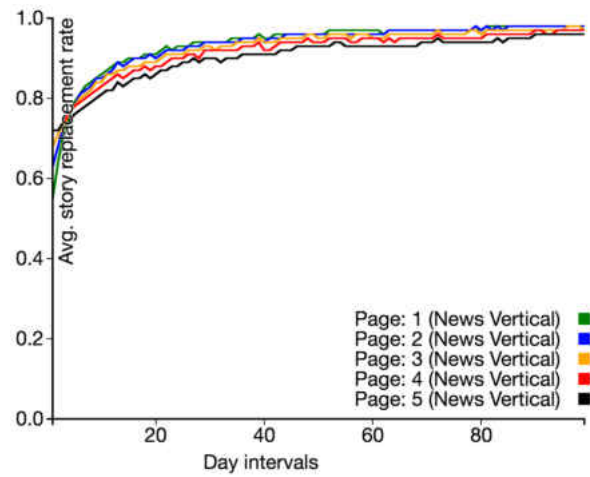
Collection	General SERP			News vertical SERP		
	Daily	Weekly	Monthly	Daily	Weekly	Monthly
healthcare bill	0.42	0.58	0.62	0.44	0.70	0.82
manchester bombing	0.27	0.37⁻	0.46⁻	0.31⁻	0.52⁻	0.66⁻
london terrorism	0.34	0.40	0.51	0.43	0.65	0.84
trump russia	0.54⁺	0.78⁺	0.83⁺	0.42	0.70	0.83
travel ban	0.43	0.62	0.71	0.45	0.61	0.75
hurricane harvey	0.21⁻	0.38	0.51	0.49	0.76	0.82
hurricane irma	0.27	0.41	0.61	0.57⁺	0.81⁺	0.91⁺

but at much faster rates. These results show us that the timing of collection building efforts that utilize SERPs is critical especially for rapidly evolving events with undefined timeframes. Since these events produce newer stories continuously, collection building must be continuous in order to capture the various cycles of the event.

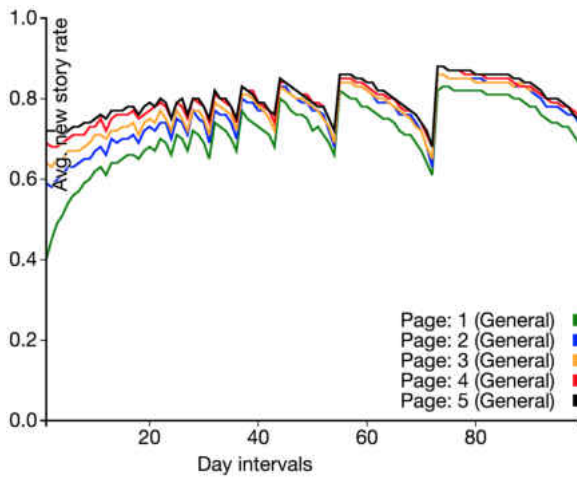
Figures 25a & 25c show that the average story replacement rate and average new story rate differed across various pages for the *General* SERP. There was a direct relationship between page number and story replacement rate (or new story rate) - the higher the page number, the higher the story replacement rate (or new story rate), and vice versa. The direct relationship may be due to fact that higher order pages (e.g., pages 4 and 5) are more likely to receive documents from lower order pages (e.g, page 1–3) than the opposite. For example, the probability of going from page 1 to page 5 was 0.0239 while the probability of going from page 5 to page 1 was 0.0048. The lower order pages have the highest quality on the SERP, thus, there is high competition within documents to retain their position on a lower order page (high rank). The competition in the higher order pages is less, therefore, when documents from the lower order pages lose some rank, they may fall into the higher order pages thereby increasing the new story rate of higher order pages. The *News* vertical SERP showed an inverse relationship between the page number and the story replacement rate (or new story rate) (Figure 25b & d) even though the probability of going from a page 1 to page 5 (0.0801) was more likely than the opposite (0.0009). This may be due to some unseen mechanism in the *News* vertical SERP.



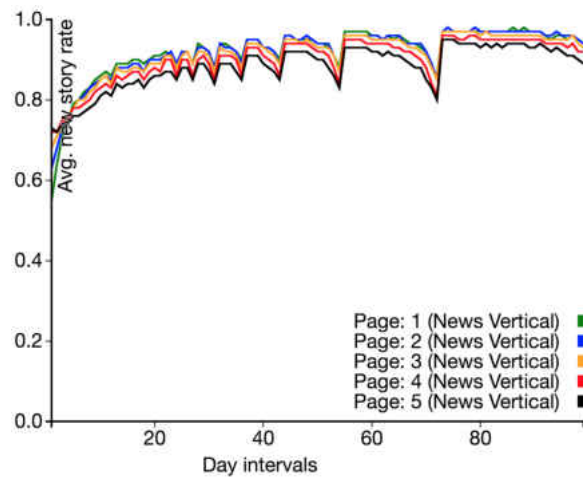
(a) The page-level average story replacement rates for *General* SERP collections show a direct relationship between page number and story replacement rate - the higher the page number, the higher the story replacement rate, and vice versa.



(b) The page-level average story replacement rates for *News* vertical SERP collections show an inverse relationship between page number and story replacement rate - the higher the page number, the lower the story replacement rate, and vice versa.



(c) Similar to the page-level average story replacement rate, the page-level average new story rate for *General* SERP collections show a direct relationship between page number and new story rate.



(d) Similar to the page-level average story replacement rate, the page-level average new story rate for *News* vertical SERP collections show an inverse relationship between page number and new story rate.

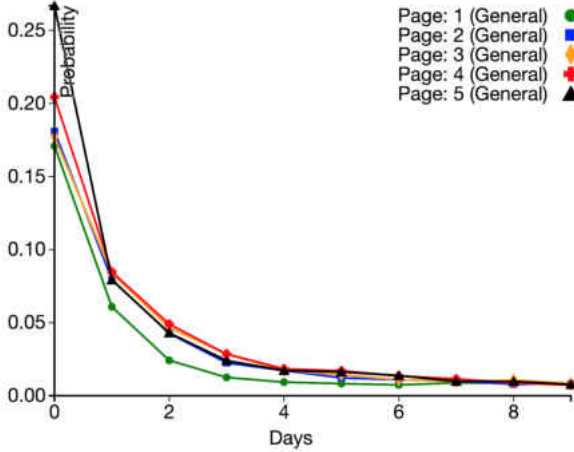
Fig. 25: a & b: Page-level new story rates for *General* and *News* vertical SERPs. c & d: Page-level story replacement rates for *General* and *News* vertical SERPs.

TABLE 11: Probability of finding the same story after one day, one week, and one month (from first observation) for *General* and *News* vertical SERP collections. Column markers: **minimum⁻** and **maximum⁺**.

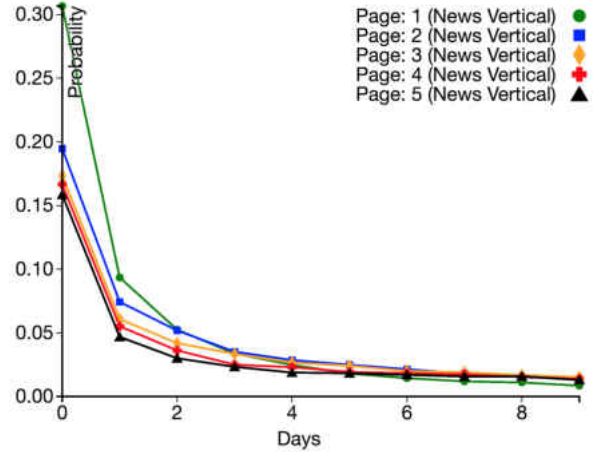
Collection	General SERP			News vertical SERP		
	a day	a week	a month	a day	a week	a month
healthcare bill	0.35	0.04	0.02	0.34	0.07	0.00
manchester bombing	0.44⁺	0.09	0.07	0.40⁺	0.14⁺	0.00
london terrorism	0.37	0.11⁺	0.07	0.34	0.09	0.00
trump russia	0.39	0.01⁻	0.01⁻	0.36	0.10	0.00
travel ban	0.43	0.06	0.02	0.32	0.12	0.00
hurricane harvey	0.38	0.10	0.08⁺	0.29	0.05	0.00
hurricane irma	0.34⁻	0.07	0.05	0.28⁻	0.03⁻	0.00

4.2.2 PROBABILITY OF FINDING A STORY

Table 11 shows the probability of finding the same story after one day, one week, and one month (from first observation) for *General* and *News* vertical SERP collections. The probability of finding the same URI of a news story with the same query decreased with time for both SERP collections. For the *General* SERP, the probability of the event that a given URI for a news story is observed on the SERP when the same query is issued one day after it was first observed ranged from 0.34 – 0.44. When the query was issued one week after, the probability dropped to from 0.01 – 0.11, one month after - 0.01 – 0.08. The probability of finding the same story with time is related to the rate of new stories: for a given time interval, the higher the rate of new stories, the lower the chance of observing the same story, because it is more likely to be replaced by another story. For example, compared to the *manchester bombing* collection, the *hurricane irma* collection produced a lower (0.34) probability (vs. *manchester bombing* - 0.44) of finding the same story after one day due to its higher (0.79) new story rate after one day (vs. *manchester bombing* - 0.52). The probability of observing the same news story on the *News* vertical SERP declined with time, but at a much faster rate compared to the *General* SERP. In fact, Table 11 shows that for all seven topics in the dataset, the probability of finding the same story on the *News* vertical when the query was re-issued one month after was marginal (approximately 0.0). This is partly because the *News* vertical SERP collections produced higher story replacement and new



(a) Probability of finding a story after variable number of days on pages (1 – 5) for *General* SERP shows direct relationship between page number and probability



(b) Probability of finding a story after variable number of days on pages (1 – 5) for *News* vertical SERP shows inverse relationship between page number and probability

Fig. 26: a & b: Page-level probability of finding the URI of a story over time.

story rates than the *General* SERP collections.

In order to generalize the probability of finding an arbitrary URI as a function of time (days), we fitted a curve (Figure 27) over the union of occurrence of the URIs in our dataset with an exponential model. The probability $P_{s,sp}(k)$ of finding an arbitrary URI of a news story s on a SERP $sp \in \{General, NewsVertical\}$, after k days is predicted as follows:

$$P_{s,General}(k) = 0.0362 + 0.9560e^{-0.9159k}$$

$$P_{s,NewsVertical}(k) = 0.0469 + 0.9370e^{-0.9806k}$$

$P_{s,sp}(k)$ predicts the probability of finding an arbitrary URI while $P(s^{dk})$ (Equation 3) produces an empirical probability of finding an arbitrary URI. Also, similar to the story replacement and new story rates, for the *General* SERP, the results showed a direct relationship with the page number and probability of finding news stories over time (Figure 26a). For the *General* SERP, higher order page numbers (e.g., 4 and 5) produced higher probabilities of finding the same stories compared to lower order (e.g., 1 and 2) pages. This might be because during the lifetime of a story, the probability of the story going from a lower order (high rank) page to a higher (low rank) order page is higher than the opposite - going from higher order page to lower order page (climbing in rank). For example, the probability of going from page 1 to page 5 was higher (0.0239) than the probability of going

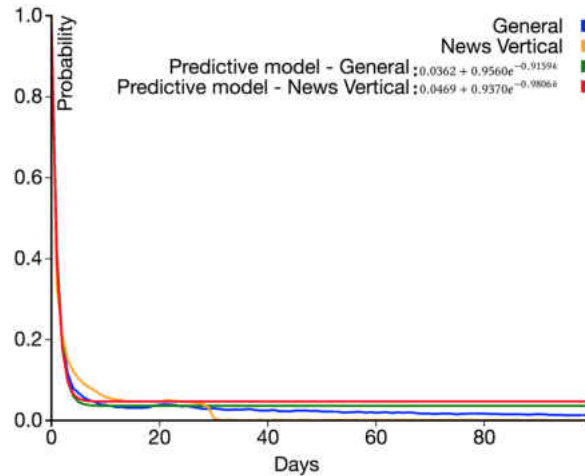


Fig. 27: Probability of finding an arbitrary story for *General* and *News* vertical SERPs was modeled with two best-fit exponential functions. In general, the probability of finding the URI of a news story on the *General* SERP is higher (lower new story rate) than the probability of finding the same URI on the *News* vertical SERP (due to its higher new story rate).

from page 5 to page 1 (0.0048). However, collections from *News* verticals showed that the lower the page number, the higher the probability of finding news stories (inverse relationship) even though the probability of falling in rank (lower order page to higher order page) is higher than the probability of climbing in rank (higher order page to lower order page).

4.2.3 DISTRIBUTION OF STORIES OVER TIME ACROSS PAGES

Figure 28 shows how the temporal distributions typically differ between *General* and *News* vertical SERP collections. There are two dimensions in the figure: days (x-axis) and URIs of stories (y-axis). A single dot in the figure indicates that a specific story occurred at that point. The temporal distribution is a reflection of the new story rate, but at a granular (individual) story level. *General* SERP collections had lower new story rates, thus produced stories with a longer lifespan than *News* vertical SERP collections. In Figure 28, this is represented by a long trail of dots. Since *News* vertical collections had higher story replacement and new story rates, they produced documents with shorter lifespans. For example, Figure 28a contrasts the denser (longer lifespan) temporal distribution of the “hurricane harvey” *General* SERP collection to the sparser “trump russia” *General* SERP collection (Figure 28c). The “trump russia” collection produced new documents on

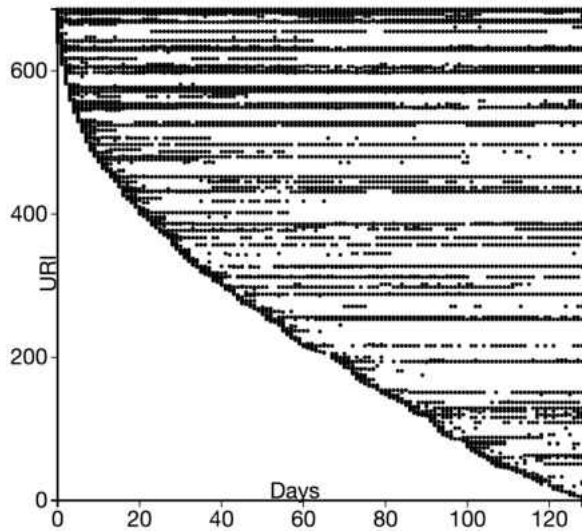
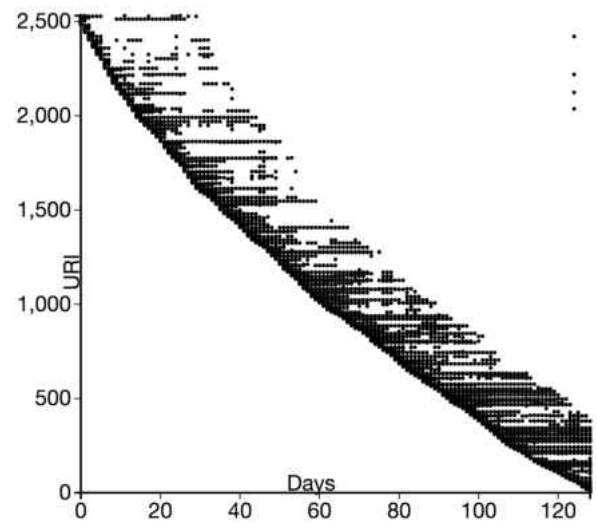
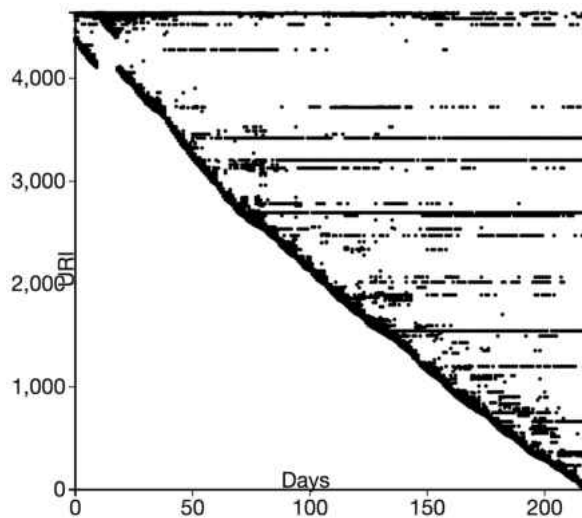
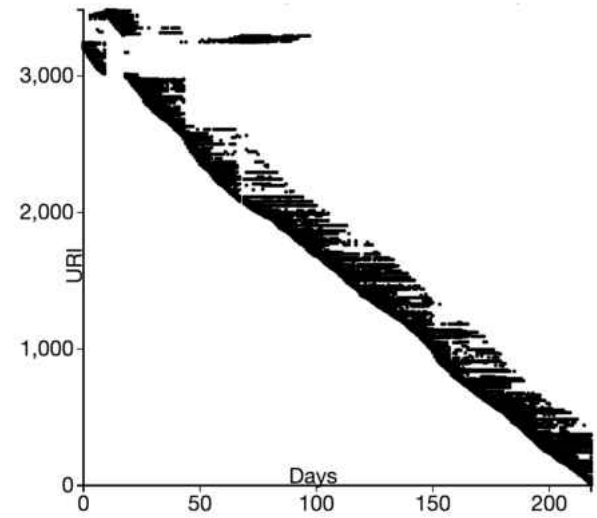
(a) “hurricane harvey” *General* SERP collection(b) “hurricane harvey” *News* vertical SERP collection(c) “trump russia” *General* SERP collection(d) “trump russia” *News* vertical SERP collection

Fig. 28: Temporal distributions: Stories in General SERP collections (a & c) persist longer (“longer life”) than stories in *News* vertical collections (b & d). Compared to the “trump russia” *General* SERP collection, the stories in the “hurricane harvey” *News* vertical collection have a “longer life” due to a lower rate of new stories.

average at a rate of 0.54 (daily) to 0.83 (monthly), compared to the “hurricane harvey” collection (daily - 0.21, and monthly - 0.51). Similarly, since documents from the “trump russia” collections were rapidly replaced (story replacement rate: 0.54 – 0.92) with newer documents, they mostly did not persist on the SERP.

Figures 29 and 30 show how URIs moved across pages over time. The rows represent the URIs and the columns represent the pages in which the URIs were observed on a specific day. A single cell represents the page in which a URI occurred on a specific day. For example, the first cell (row 0, column 0) of Figure 29 is 1. This means the URI at row 0 was first observed on page 1. Some of the same URIs persist over time within the same page. For example Figure 29, row 0, shows that the highly ranked Wikipedia page² of the *Manchester bombing* event was seen for 24 consecutive days on the first page of the SERP, was not seen (within page 1 – 5) on the 25th day, and then seen for 13 consecutive days (still on page 1). Figure 29 also shows the increase/decrease in ranks for stories. For example, in Figure 29, row 4, the URI³ was first observed on page 5, the next day it increased in rank to page 1, skipping 2 – 4. The page-level temporal distribution also shows that some stories go directly from page 5 to 1. In contrast with *General* SERP collections, the temporal distribution of *News* vertical collections is shorter (Figure 30) and reflect the higher story replacement and new story rates of *News* vertical collections.

4.2.4 OVERLAP AND RECALL

Table 12 shows that setting the Google date range parameter improves finding stories with respect to the set date range for both *General* and *News* vertical collection. For example, for the “healthcare bill” *General* SERP collection, the *Jan-2018* collection which was created (2018-01-11) by making a default search (without) setting the date range had an overlap rate of 0.06 with respect to the collection of documents created in June 2017 (*June-2017*). In contrast, the collection created the same day (2018-01-11) by setting the date range parameter to June 2017 (2017-06-01 to 2017-06-30) had a much higher overlap rate of 0.60. This is the case across all collection topics, especially for topics with lower new story rates (0.27 - 0.46) such as “manchester bombing” (0.82 overlap rate). The *News* vertical collections had lower overlap rates compared to the *General* SERP collections since *News* vertical collection have higher story replacement and new story rates.

Irrespective of the increase in refinding (overlap) new stories that occurs when the date

²https://en.wikipedia.org/wiki/Manchester_Arena_bombing

³<http://www.dailymail.co.uk/news/article-4578566/Evidence-Nissan-linked-Manchester-bombing.html>

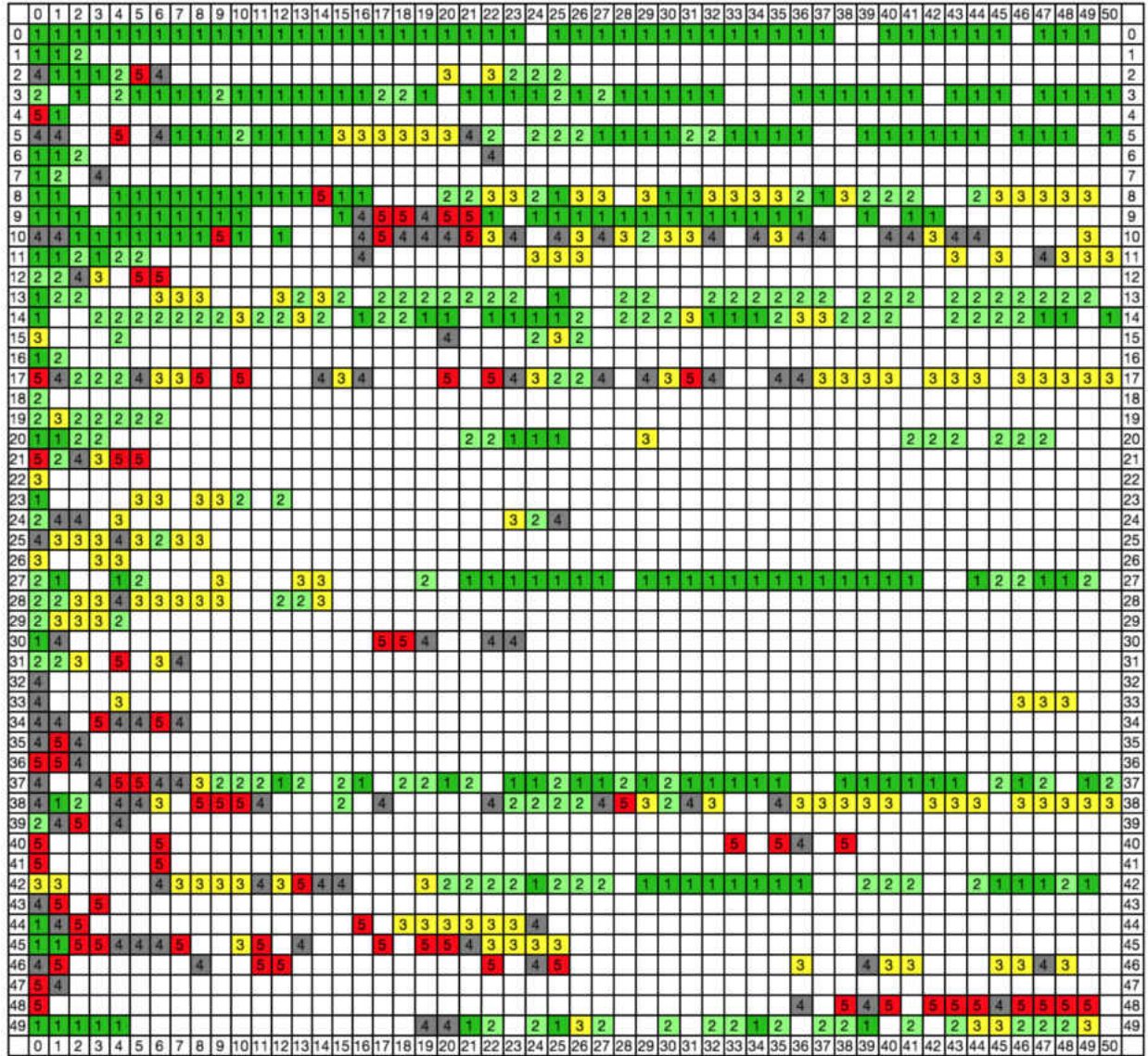


Fig. 29: Page-level temporal distribution of stories in the “manchester bombing” *General* SERP collection showing multiple page movement patterns. Stories in *General* SERP collections persist longer than stories in *News* vertical collections. Color codes - page 1, page 2, page 3, page 4, page 5, and blank for outside pages 1 – 5.

TABLE 12: Comparison of two collections against the *June-2017* collection (documents published in June 2017). The collection *Jan-2018*, which was created (2018-01-11) without modifying the SERP date range parameter has a lower overlap than the collection (*June-2018-Restricted-to-June*) created the same day (2018-01-11) by setting the SERP date range parameter to June 2017. Even though setting the date range parameter increases finding stories with common publication dates as the date range, the recall is poor due to the fixed SERP result. Column markers: **maximum**.

Collection	Metrics	General SERP			News vertical SERP		
		<i>June-2017</i>	<i>Jan-2018</i>	<i>Jan-2018-Restricted-to-June</i>	<i>June-2017</i>	<i>Jan-2018</i>	<i>Jan-2018-Restricted-to-June</i>
healthcare bill	size	460	51	50	419	50	50
	overlap	1.00	0.06	0.60	1.00	0.02	0.56
	recall	1.00	0.01	0.07	1.00	0.00	0.07
manchester bombing	size	483	50	51	50	50	548
	overlap	1.00	0.04	0.82	1.00	0.00	0.50
	recall	1.00	0.00	0.08	1.00	0.00	0.05
london terrorism	size	191	50	52	50	50	172
	overlap	1.00	0.09	0.70	1.00	0.00	0.68
	recall	1.00	0.02	0.18	1.00	0.00	0.20
trump russia	size	562	50	51	50	50	524
	overlap	1.00	0.00	0.54	1.00	0.00	0.58
	recall	1.00	0.00	0.05	1.00	0.00	0.06
travel ban	size	391	50	52	50	50	370
	overlap	1.00	0.04	0.84	1.00	0.16	0.48
	recall	1.00	0.01	0.11	1.00	0.02	0.06

range parameter is set, the recall is poor. Since the SERP only produces a fixed number of documents per page, we only get a small fraction of the documents relevant to the specified date range. The “healthcare bill” *June-2017 General* SERP collection contains 460 documents published in June 2017, collected by extracting URIs from the first five pages of the SERP. A query (“healthcare bill”) issued to the SERP in January 2018, with the date range parameter set to June 2017 increased overlap (refinding stories), but did not increase the number of results - we could only extract at most approximately 50 URIs (first five pages). Consequently, across all topics in Table 12, both *Jan-2018* and *Jan-2018-Restricted-to-June* collections had recall of under 0.10 except for the “london terrorism” topic (maximum recall 0.20). This reaffirms the idea that collection building or seed selection processes that rely on the SERP must start early and persist in order to maximize recall. To further aid selection of seeds, a simple set of heuristics could identify most of the likely stable URIs (e.g., *wikipedia.org*, *nasa.gov*, *whitehouse.gov*) as well as URIs likely to quickly disappear from the top-k SERPs (e.g., *cnn.com* or *nytimes.com*, followed by a long path in the URI). The archivist could give priority to the latter URIs, knowing that the former URIs will continue to be discoverable via Google.

4.3 GENERATING SEEDS FROM SERPS, A RECOMMENDATION

Search engines provide an opportunity to extract seeds, but tend to provide the most recent documents. Our findings illustrate the difficulty in refinding news stories as time progresses. On average, the rate at which stories were replaced on the Google *General* SERP ranged from 0.21 – 0.54 daily, 0.39 – 0.79 weekly, and 0.59 – 0.92 monthly. The Google *News* vertical SERP showed even higher story replacement rates, with a range of 0.31 – 0.57 daily, 0.54 – 0.82 weekly, and 0.76 – 0.92 monthly. Also, the probability of finding the same news story diminishes with time and is query dependent. The probability of finding the same news story with the same query again, one day after the first time the story was first seen ranged from 0.34 – 0.44. If one waited a week, or a month and issued the same query again, the probability of finding the same news story drops to 0.01 – 0.11. The probability declines even further if we used the *News* vertical SERP due to its higher story replacement and new story rates. Discoverability may be improved by instructing the search engine to return documents published within a temporal range, but this information is not readily available for many events, and we discover only a small fraction of relevant documents since the count of search results are restricted.

The web archiving community considers link rot and content drift important reasons for

collection building. Similarly, our findings suggest that due to the difficulty in retrieving the URIs of news stories from Google, collection building that originates from search engines should begin as soon as possible in order to capture the first stages of events, and should persist in order to capture the evolution of the events, because it becomes more difficult to find the same news stories with the same queries on Google, as time progresses. The *SERP-Refind* dataset comprising of 151,602 (33,432 unique) links extracted from the Google SERPs for over seven months, as well as the source code for the application utilized to semi-automatically generate the collections, are publicly available [141].

4.4 CHAPTER SUMMARY

SERPs are a popular source for seeds but are known to produce the most recent URIs corresponding to the time a query is issued. This chapter provided a study to quantify this phenomenon and to understand how difficult it is to refind the URIs of news stories on SERPs as a function of time. We discovered that due to the high story replacement rates of the Google SERPs, it is improbable to find the URI of a news story with the same query after one week. It is highly improbable to find the same URI when the query is issued after one month. Additionally, the improbability of finding URIs of news stories increases when considering long-running ongoing topics. These findings collectively express the difficulty in refinding news stories with time, thus motivates the need for collection building processes that utilize the SERP to begin early and persist in order to capture the start and evolution of an event.

CHAPTER 5

SCRAPING SEEDS FROM MICRO-COLLECTIONS IN SOCIAL MEDIA

Two main strategies adopted by curators for discovering seeds include scraping Web and social media SERPs. In Chapter 4 we explored collections generated from SERPs. Here, we shift our focus to those generated from social media, and we address the first two research questions (Chapter 1.4) of this effort, repeated here for convenience:

- **RESEARCH QUESTION 1:** How do we identify, extract, and profile Micro-collections in social media?
- **RESEARCH QUESTION 2:** Do seeds from Micro-collections differ from seeds from SERPs and hashtags?

This chapter presents the findings and results of our study [145] of three social media platforms (Reddit, Twitter, and Scoop.it) to address the first two research questions. We begin by presenting a vocabulary (*post class*, Chapter 5.1) for labeling social media posts across different platforms. Next, we identify, extract, and characterize Micro-collections in social media and show how their seeds differ (Chapter 5.4) from those generated from conventional sources such as Web (e.g., Google) and social media (e.g., Twitter) SERPs for text and hashtag queries. The differences are potentially consequential to curators generating seeds from social media with specific needs, therefore the study concludes by presenting a recommendation for generating seeds from social media (Chapter 5.5).

5.1 POST CLASS: CLASSIFICATION SYSTEM FOR LABELING SOCIAL MEDIA POSTS

Before extracting and studying seeds from social media, it was essential to define a means of labeling the various kinds of social media posts. Since we extracted seeds from multiple social media platforms, it was also essential to provide generic labels for social media posts. A generic label for social media posts regardless of platform enables discussion about posts from different social media sites such as Reddit and Twitter. Consequently, we developed

the *post class* (Table 13) system of labeling social media posts regardless of platform. The post class ($\mathbf{P}_*\mathbf{A}_*$) consists of four members ($\mathbf{P}_1\mathbf{A}_1$, $\mathbf{P}_1\mathbf{A}_n$, $\mathbf{P}_n\mathbf{A}_1$, $\mathbf{P}_n\mathbf{A}_n$) that are pairs of acronyms that identify social media posts regardless of platform. A single post class is formed by combining two acronyms, \mathbf{P} and \mathbf{A} , with subscripts (1 - single or n - multiple), both combined to represent the count of \mathbf{P} osts and \mathbf{A} uthors, respectively.

TABLE 13: Post class for social media posts. All non- $\mathbf{P}_1\mathbf{A}_1$ collections are combined to create Micro-Collections (MC). However, some $\mathbf{P}_1\mathbf{A}_1$ posts (e.g., Figure 32) can be considered as Micro-collections if they contain more links than the median number of links estimated for $\mathbf{P}_1\mathbf{A}_1$ posts of the social media platform.

Post Class	Post Count	Author Count	Definition/Example
$\mathbf{P}_1\mathbf{A}_1$	Single (1)	Single (1)	A Single P ost from a single A uthor, e.g., an isolated tweet or Reddit post (Figure 32a & 32b). These posts are visible to seeds generators that scrape SERPs.
$\mathbf{P}_1\mathbf{A}_n$	Single (1)	Multiple (n)	Single P ost from multiple A uthors, e.g., the references contributed by multiple Wikipedia editors.
$\mathbf{P}_n\mathbf{A}_1$	Multiple (n)	Single (1)	Multiple P osts from a single A uthor, e.g., a thread of tweets (Figure 31) from a Twitter user.
$\mathbf{P}_n\mathbf{A}_n$	Multiple (n)	Multiple (n)	Multiple P osts from multiple A uthors, e.g., a tweet conversation consisting of multiple tweets or posts from different Twitter or Reddit (or Facebook) users.

We introduced Micro-collections in Chapter 1, and we repeat examples (Figures 31 and 32) here for convenience. Recall that a Micro-collection refers to social media posts authored by single or multiple authors that exhibit certain properties associated with collection building. For example, the Twitter account *Doing Things Differently* (@dtdchange) [149] created a chain of tweets by replying to each subsequent tweet in order to chronicle the *Flint Water Crisis* story. This reply thread spans almost 3 years and consists of 75 tweets (as of January 9, 2019) each containing a URI. It is fair to attribute curatorial discretion (selection and filtering) to this collection of tweets, thus we consider it a Micro-collection for the *Flint Water Crisis* story. Another example of Micro-collections are Reddit posts created by the user *Ilensine* [150] for the 2014 *Ebola virus outbreak* story. In total, the posts contain over 102 external references and were published less than two weeks after the World Health Organization (WHO) declared the 2014 Ebola outbreak a Public Health Emergency of International Concern [151]. We distinguish Micro-collections from standard social media posts by showing that some kinds of Micro-collections can be identified by considering the

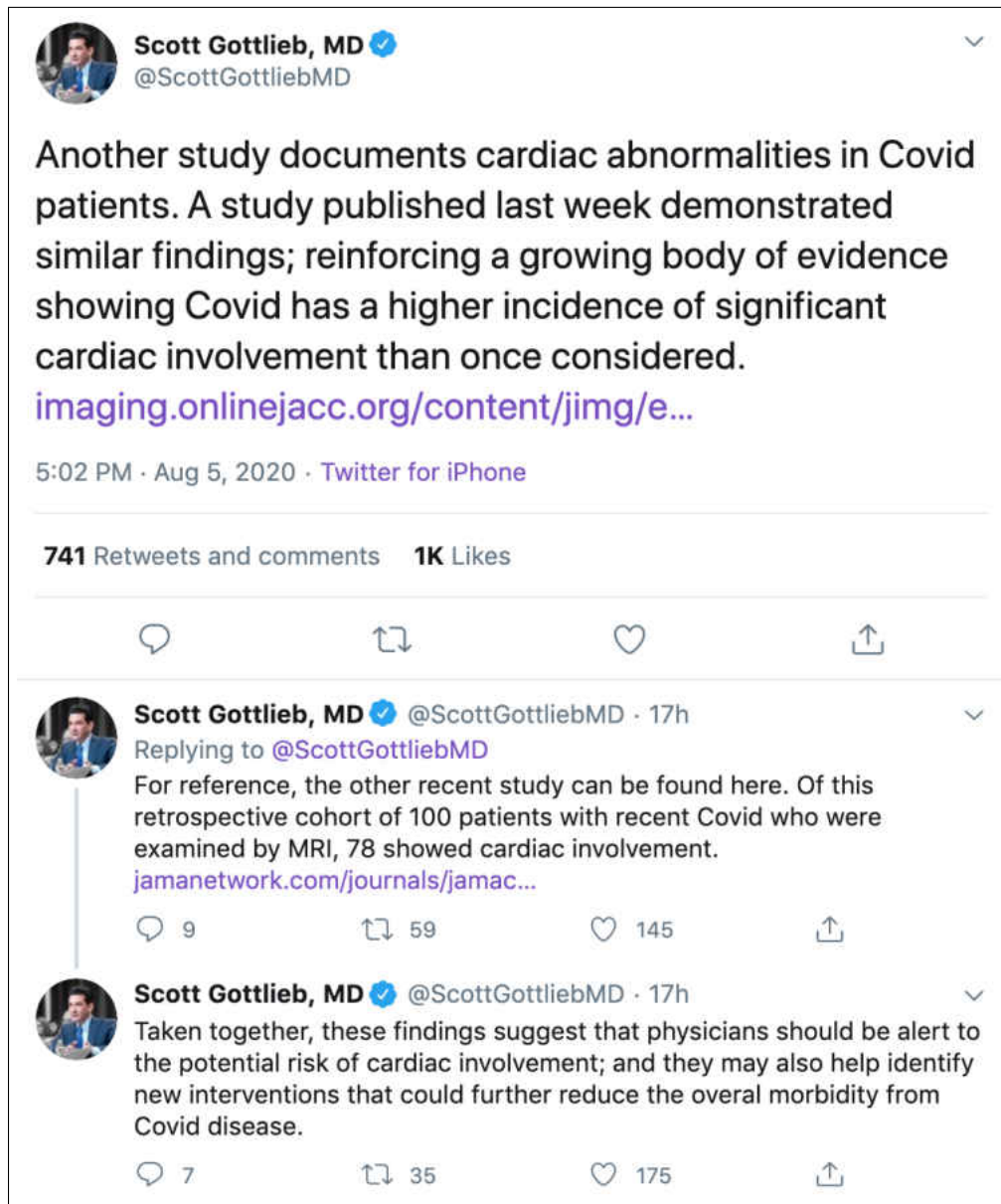


Fig. 31: Example of a Micro-collection from Twitter by a single author (@ScottGottliebMD) consisting of three tweets that are part of a reply thread [146] about the *2020 Coronavirus Pandemic*. This Micro-collection is of post class $\mathbf{P}_n\mathbf{A}_1$ since it consists of multiple Posts from a single Author. This image has been edited to show more details.

Posted by u/Isensine 4 years ago 🟡 🟢

[Ebola] 2014 Outbreak Report

Main Report

Please see [PART 2](#) for continuing daily updates.
This post is now an Archive for the dates below. I have exceeded the character limit, but makes perfect sense looking at it now.

[Ebola Hemorrhagic Fever](#) (EBOV) is one of numerous Viral Hemorrhagic fevers that can affect humans and nonhuman primates

The outbreak started in West Africa and is from the [Zaire Strain](#).
[The WHO entry](#)
[The CDC entry](#).

- [Ebola Infographic](#)
- [The Wikipedia entry](#).
- [HealthMap - Ebola](#)
- [Interactive Outbreak Tracker](#)

[2014 - Aug - 19] (<http://redd.it/2dze0y>)

- Images during Liberia's Ebola outbreak (editorial).
- Surviving Ebola, but untouchable back home (editorial).
- Situation overviews from Al Jazeera and Reuters (editorial).
- Liberian boarder guards ordered to Shoot-on-sight anyone crossing the border.
- Liberian government orders curfew over Ebola outbreak.
- Missing Ebola patients found (Liberia).
- WHO has requested all affected countries conduct exit screening.
- WHO: Ebola situation in Nigeria and Guinea: encouraging signs.
- 600 people quarantined for hours over Ebola scare, turned out to be a false alarm.

(a) Part 1: Reddit post

Posted by u/Isensine 4 years ago 🟡 🟢

Ebola 2014 Outbreak Report: pt2

Main Report

This is Part Two, I have moved all of the daily updating info here.

Primary Ebola Strain:

Updated: 2014 - November 10th

Total Cases as of November 4th [\[WHO\]](#)
Suspected and Confirmed Case Count: 13,268
Suspected and Confirmed Case Deaths: 4,960

[Guinea](#)
Suspected and Confirmed Case Count: 1,760
Suspected Case Deaths: 1054
Percent of Population (11,474,383) Affected: 0.0134%

[Liberia](#)
Suspected and Confirmed Case Count: 6,619
Suspected Case Deaths: 2,766
Percent of Population (4,092,310) Affected: 0.1140%

[Sierra Leone](#)
Suspected and Confirmed Case Count: 4,862
Suspected Case Deaths: 1,130
Percent of Population (5,743,725) Affected: 0.0645%

Nigerian Outbreak Declared Contained.

Last Update: 2014 - October - 20th [\[1\]](#)
Suspected and Confirmed Case Count: 20
Suspected Case Deaths: 8
Percent of Population (177,155,754) Affected: 0%

(b) Part 2: Reddit post

Fig. 32: Example of a pair of Micro-collection Reddit posts [147, 148] consisting of 102 external references for the 2014 Ebola outbreak. Both Micro-collections are of type $\mathbf{P}_1\mathbf{A}_1$ (single Posts from a single Author).

properties of the posts.

5.2 EXPERIMENT: CHARACTERIZING AND COMPARING SERP AND MICRO-COLLECTION SEEDS

To enable characterizing and comparing collections, first we selected five topics (Chapter 5.2.1) and generated a dataset from the topics consisting of seeds extracted from social media posts. Next, we segmented (Chapter 5.2.2) the social media posts (and the seeds they contain) of the dataset into their respective post class described in Chapter 5.1. A collection is simply a list of seeds (URIs) extracted from social media posts that belong specific kind of post class (e.g., Micro-collections). Additionally, we generated a gold standard to enable assessing the precision of the seed collections. These enumerated steps enabled comparing and characterizing seed collections.

5.2.1 TOPIC SELECTION

A central objective of the discussed research was to outline the characteristics of, and differences between, collections generated by scraping SERPs ($\mathbf{P}_1\mathbf{A}_1$ post class - Table 13) and Micro-collections (\mathbf{MC} post class). Therefore, the choice of queries was not arbitrary. Instead, to gain an approximate representative dataset sample to study, we developed a temporal classification system (partly informed by Gossen et al. [110]) of real world stories and events based on three temporal (Table 14) attributes: *Expectation* (event expected or unexpected), *Recurrence* (recurring or non-recurring event), and *Occurrence definition* (start and end times defined or undefined). A story can be described by a combination of different states of the temporal attributes.

For the expectation attribute, an event may be *expected* or *unexpected*. For example, the *Ebola outbreak* event was unexpected. Thus we classify this event as an unexpected event. For the recurrence attribute, an event may occur repeatedly at regular or non-regular intervals. For example, the FIFA World Cup tournaments are played at four-year intervals, thus we consider this event a recurring event. Ebola outbreaks in general may also be considered a recurring event, even though they occur at irregular intervals. For the occurrence definition attribute, an event may have a defined or undefined start and end date. For example, the *MSD Shooting* event started and ended the same day (February 14, 2018), but the *Flint Water Crisis* event started in April 2014, and is still ongoing (no end definition).

TABLE 14: Temporal characteristics of the Micro-collections dataset topics

Topic [Wikipedia Page]	Expectation (Expected/ Unexpected)	Recurrence (Recurring/ Non-Recurring)	Occurrence definition	
			Start (Defined/ Undefined)	End (Defined/ Undefined)
Ebola Virus Outbreak [152]	Unexpected	Recurring (Irregular)	Dec 2013 [153]	Jun 2016 [153]
Flint Water Crisis [154]	Unexpected	Non-Recurring	Mar 2014 [31]	Undefined
MSD Shooting [155]	Unexpected	Non-Recurring	Feb 14, 2018	Feb 14, 2018
2018 World Cup [156]	Expected	Recurring	Jun 14, 2018	Jul 15, 2018
2018 Midterm Elections [157]	Expected	Recurring	Nov 6, 2018	Nov 6, 2018

Following the specification of the temporal classification system, we selected five topics specified by the following queries and hashtags (for Twitter): “ebola virus outbreak” (#ebolavirus), “flint water crisis” (#FlintWater), “stoneman douglas high school shooting” (#MSDStrong), “2018 world cup” (#WorldCup) and “2018 midterm elections” (#election2018). Table 14 presents the temporal attributes of each of these selected topics. In addition to text queries, for Twitter, we selected hashtag queries for each topic to discern if seeds generated with text-based queries differ from those extracted with hashtag queries.









5.2.2 DATASET GENERATION AND SEGMENTATION OF SOCIAL MEDIA POSTS INTO POST CLASSES

For Reddit, we issued all five queries to four Reddit SERPs (Relevance, Top, New, and Comments), and extracted posts from the SERPs. For each query we extracted a maximum of 500 posts and recursively extracted a maximum of 500 comment replies from each post extracted from the SERP.

For Twitter, similar to Reddit, we issued all five text and hashtag queries to the two Twitter SERPs (Top and Latest), and extracted tweets from the SERPs with the use of the Local Memory Project’s [142] *local news generator* [143]. For each query, we extracted a maximum of 500 tweets and recursively extracted a maximum of 500 tweet replies for each tweet extracted from the SERP.

For Reddit and Twitter, the posts directly visible from the SERP were assigned to the $\mathbf{P}_1\mathbf{A}_1$ post class (Chapter 5.1 and Table 13). We use the term “post” in order to be general. Different social media sites have different names for posts, for example, on Twitter, a post is called a tweet. Posts with replies were assigned either to the $\mathbf{P}_n\mathbf{A}_1$ or $\mathbf{P}_n\mathbf{A}_n$ class depending

TABLE 15: Post class counts (Class), Social media posts (Posts), and URI counts (URIs) for dataset generated by extracting URIs from post classes ($\mathbf{P}_1\mathbf{A}_1$, $\mathbf{P}_n\mathbf{A}_1$, and $\mathbf{P}_n\mathbf{A}_n$) of Reddit, Twitter, Twitter Moments, and Scoop.it. The Micro-collection (MC) post class is formed by combining posts in $\mathbf{P}_n\mathbf{A}_1$ and $\mathbf{P}_n\mathbf{A}_n$ post classes.

	Micro-collections (MC)								
	$\mathbf{P}_1\mathbf{A}_1$ Counts			$\mathbf{P}_n\mathbf{A}_1$ Counts			$\mathbf{P}_n\mathbf{A}_n$ Counts		
	Class	Posts	URIs	Class	Posts	URIs	Class	Posts	URIs
Reddit Relevance 	766	766	1,776	56	115	206	542	36,124	3,387
Reddit Top 	931	931	10,857	37	177	319	1,021	100,006	18,992
Reddit New 	854	854	8,056	26	68	1,062	340	9,298	6,412
Reddit Comments 	834	834	8,381	53	423	691	1,077	117,378	18,781
Twitter Top 	2,936	2,936	3,548	540	4,983	3,026	4,009	79,347	12,457
Twitter Latest 	2,341	2,341	2,792	639	6,366	3,628	4,471	82,499	13,576
Twitter Moments 	NA	NA	NA	NA	NA	NA	73	1,285	621
Scoop.it 	1,533	1,533	1,533	33	1,083	343	NA	NA	NA
Subtotal	10,195	10,195	36,943	1,384	13,215	9,275	11,533	425,937	74,226
Total	Class: 23,112			Posts: 449,347			URIs: 120,444		

on the number of authors. Posts from the SERP with a reply or a contiguous set of replies exclusively authored by a single user were assigned to the $\mathbf{P}_n\mathbf{A}_1$ post class. Finally, posts with a reply or a series of replies authored by multiple users were assigned to the $\mathbf{P}_n\mathbf{A}_n$ post class. The $\mathbf{P}_1\mathbf{A}_n$ Micro-collection post class is rare and not available in Twitter, Reddit, or Scoop.it. However, our gold standard data was extracted from Wikipedia references which belong to $\mathbf{P}_1\mathbf{A}_n$.

For Twitter Moments, we issued all five queries to Google with “`site:twitter.com/i/moments`” in order to restrict the search results to links from Twitter Moments. Next, we extracted Twitter Moments URIs from the first two pages of the Google default SERP. Next, we dereferenced URIs and extracted the tweets. Tweets from Twitter Moments are authored by multiple users, and thus assigned the $\mathbf{P}_n\mathbf{A}_n$ label.

In addition to the extraction of posts from well-known social media (Reddit and Twitter), we considered a lesser known social media site called Scoop.it (<https://www.scoop.it/>). Scoop.it is a content curation social media service that enables users to bookmark a single URI (*scoop*) or multiple URIs (*topics*). For Scoop.it, we issued all five queries to the Scoop.it SERPs (Scoops and Topics), and extracted posts (scoops) from the SERPs. The scoops visible from the *Scoops* SERP were assigned to the $\mathbf{P}_1\mathbf{A}_1$ post class. For a single dataset topic, the scoops found in the *Topic* SERP were assigned to the $\mathbf{P}_n\mathbf{A}_n$ post class since they are authored by multiple users.

From all social media posts, we extracted the URIs to create collections corresponding to the post class from which the URIs were extracted. Social media posts often link to intra-site posts (e.g., tweet URI in a tweet). We dereferenced and extracted seeds from such intra-site URIs, and substituted them with the extracted seeds.

5.2.3 GOLD STANDARD DATASET GENERATION

The following steps were taken in order to generate the gold standard dataset to facilitate measuring precision of URI collections extracted from the various post classes. First, we selected a corresponding Wikipedia page for the five topics (Table 14). Second, we extracted the URIs from the references section of each Wikipedia page. Third, we dereferenced the URIs from each reference corresponding to a topic (e.g., *Flint Water crisis*) and removed the HTML boilerplate leaving only the plaintext documents (stopwords removed). The set of plaintext documents were concatenated into one document. Fourth, for each topic, we created a collection vector consisting of the normalized Term Frequency (TF) weights of the concatenated document.

5.3 EVALUATION: METRICS FOR CHARACTERIZING AND COMPARING SERP AND MICRO-COLLECTION SEEDS

The following metrics were extracted from the Micro-collection dataset [158] to address the first two research questions.

URI and post counts per post class

We counted the number of URIs (HTML, non-HTML, and both) per topic, per social media source, and per post class (Table 15). Additionally, we extracted the distribution of posts with URIs by counting the number of posts with a specified number of links for a given social media source (e.g., Reddit) to facilitate probability distribution calculation (Table 16). The distribution answers questions such as: “for Reddit posts with links, how many posts had 1 link or 2 links?”

Probability distribution of posts with links

For all topics T (e.g., *World cup*), given the set of post classes $C \in \{\mathbf{P}_1\mathbf{A}_1, MC, \mathbf{P}_1\mathbf{A}_n, \mathbf{P}_n\mathbf{A}_1, \mathbf{P}_n\mathbf{A}_n\}$, given a social media seed source s (e.g., Reddit), the probability $P(p_c^s = k)$ of the event that a post p_c^s of post class $c \in C$ with a URI, has k URIs (e.g., 1 URI) is calculated using Equation 5. $P(p_{P_1A_1}^{Reddit} = 1)$ reads: “What is the probability of the event that a Reddit $\mathbf{P}_1\mathbf{A}_1$ post with a URI has one URI?”

The general probability $P(p_{All}^s = k)$ of the event that a post p_{All}^s with a URI from social media s of any post class, has k URIs is calculated using Equation 6. In Equation 5 & 6, if $c = P_1A_1$ and $t = 1$, $|c_1|$ represents the count of $\mathbf{P}_1\mathbf{A}_1$ posts for the first ($t = 1$) topic.

$$P(p_c^s = k) = \sum_{t=1}^{|T|} \frac{p_{c_t}^s = k}{|c_t|} \quad (5) \quad P(p_{All}^s = k) = \sum_{t=1}^{|T|} \sum_{c \in C} \frac{p_{c_t}^s = k}{|c_t|} \quad (6)$$

Precision of the URIs in post class collections

Given a candidate collection of seed URIs C to be evaluated, the URIs may be extracted from a single post ($\mathbf{P}_1\mathbf{A}_1$) or multiple posts (e.g., $\mathbf{P}_n\mathbf{A}_1$) from a social media site (e.g., Reddit). We calculated the precision of C as follows. First, the URIs in C were processed in the same manner as the gold standard (Chapter 5.2.3), i.e., dereferenced and boilerplate removed, and $|C|$ plaintext documents concatenated. Second, a document collection matrix M was created from C and its corresponding gold standard (e.g., *Flint Water Crisis* gold

standard). The first row of matrix consisted of the gold standard vector, and the second row of the matrix consisted of the vector of C (document to be evaluated). The columns represent the normalized TF weights. Third, cosine similarity was calculated between the pair of rows. If the similarity exceeded the relevance threshold of 0.25, C was declared relevant, otherwise, it was declared non-relevant. The relevance threshold was empirically determined to produce relevant results.

For a given topic (e.g., *Flint Water Crisis*) and SERP vertical (e.g., Twitter-Top), a URI or multiple URIs may be extracted from a post authored by a single ($\mathbf{P}_1\mathbf{A}_1$) or multiple ($\mathbf{P}_n\mathbf{A}_1, \mathbf{P}_n\mathbf{A}_n$) users. Each group of URIs extracted from a post has an associated precision value (Relevant URIs / Total URIs). The average precision metric for a post class (e.g., $\mathbf{P}_1\mathbf{A}_1$) is an average over all the precision value of all posts in the post class. It provides answers to questions such as: “what is the average precision of the URIs in the $\mathbf{P}_1\mathbf{A}_1$ post class?” For non-HTML URIs we evaluated precision by extracting text from the post that embedded the URI.




Age distribution of relevant webpages per post class

The distribution of ages is an aggregation of the ages of the relevant webpages in a given post class of a given social media. The age of a webpage was calculated by finding the difference between the publication date of a webpage and the date the post containing the webpage URI was retrieved. The publication dates of webpages were extracted with CarbonDate [159], which estimates the creation date of webpages based on information polled from multiple sources such as the document timestamps, web archives, backlinks, etc. Publication dates of webpages may potentially provide useful information about the kinds of events discussed. For example, the Democratic Republic of Congo in Central Africa grappled with another Ebola outbreak (2017 – 2018). Therefore, webpages published before 2017 are not expected to discuss the 2017 outbreak.

Distribution of hostname diversity per post class

Given a collection of URIs C for a given post class of a given social media, the hostname diversity [136] of C is a single value ($d \in [0, 1]$) that reports whether C consists of URIs from a single host ($d = 0.0$, e.g., `www.cnn.com`) or distinct hosts ($d = 1.0$, e.g., `www.cnn.com` and `www.foxnews.com`). It answers questions such as: “how diverse are the hosts in the $\mathbf{P}_1\mathbf{A}_1$ post class?”

TABLE 16: Probability (e.g., $P(p_{P_1A_1}^{Reddit} = 1) = \mathbf{0.63}$) of the event that a social media post from a given post class (e.g., Reddit P_1A_1) has k HTML URIs (e.g., $k = 1$).

														
k	P_1A_1	MC	P_nA_1	P_nA_n	All	P_1A_1	MC	P_nA_1	P_nA_n	All	P_1A_1	MC	P_nA_1	All
1	.63	.23	.43	.22	.37	.98	.69	.60	.70	.75	1.00	.21	.21	.97
2	.11	.12	.13	.12	.12	.02	.17	.17	.17	.14	.00	.00	.00	.00
3-4	.06	.15	.09	.15	.12	.00	.08	.11	.08	.06	.00	.12	.12	.01
5+	.20	.50	.35	.51	.39	.00	.07	.12	.06	.05	.00	.67	.67	.03

Overlap between Google collections and post class

We measured the overlap between URIs extracted from Google and URIs extracted from a combination of social media and post class. This was done in order to determine how easy it was to find the URIs scraped from social media Micro-collections. Extracting seeds from Micro-collections requires more effort than scraping Web search engine SERPs. For example, generating a collection of URIs of the P_nA_1 or P_nA_n post class requires independently dereferencing each social media post and extracting the replies from the post. Therefore, if the URIs discovered from Micro-collections are easily discoverable via a search engine such as Google, it does not justify the extra effort of extracting seeds from Micro-collections.




5.4 RESULTS

Recall the post class (Table 13) acronyms and their respective meanings and examples: P_1A_1 (e.g., a tweet) - single **P**ost from a single **A**uthor, P_1A_n (e.g., Wikipedia reference) - single **P**ost from multiple **A**uthors, P_nA_1 (e.g., twitter thread) - multiple **P**osts by a single **A**uthor, and P_nA_n (e.g., twitter conversation) - multiple **P**osts from multiple **A**uthors. In this section, results are presented with the Maximum, Median, and Minimum (MMM) notation.

To address the first research question, we identified Micro-collections ($MC = P_nA_1 \cup P_nA_n$) as the collection of social media posts that show some properties of collection building¹. Next, we extracted the P_nA_1 and P_nA_n post classes by identifying social media

¹Some P_1A_1 posts which are visible to SERP scrapers could be added to **MC** if they contain links above the median number of links, calculated from the same pool of social media posts. However, we did not make such a distinction in our study.

TABLE 17: Conditional probability (e.g., $P(\text{relevant}|p_{P_1A_1}^{\text{Reddit}} = 1) = \mathbf{0.64}$) of the event that the URIs in a social media post from a given post class (e.g., Reddit P_1A_1) are relevant, given that the post has k (e.g., $k = 1$) HTML URIs. Column markers: **minimum** and **maximum**. For the P_1A_1 , $k = 5+$ Twitter cell, the probability was calculated for just one post with eight HTML URIs.

														
k	P_1A_1	MC	P_nA_1	P_nA_n	All	P_1A_1	MC	P_nA_1	P_nA_n	All	P_1A_1	MC	P_nA_1	All
1	.64	.54	.54	.54	.60	.63	.60	.49	.61	.61	.76	.00	.00	.76
2	.80	.59	.57	.59	.65	.50	.61	.64	.60	.60	NA	.00	.00	.00
3-4	.62	.45	.50	.44	.48	.33	.46	.51	.45	.46	NA	.50	.50	.50
5+	.51	.50	.53	.50	.50	1.00	.42	.46	.41	.42	NA	.59	.59	.59

posts with replies (comments) and extracted the parent post as well as the child posts.

Following the identification and extraction of Micro-collections, to address the second research question, we characterized MCs and compared seeds extracted from them to seeds extracted from SERPs (P_1A_1). Here we present the results for each of the respective measures introduced in Chapter 4.1.2.

5.4.1 URI AND POST COUNTS PER POST CLASS

Micro-collections (MCs) are prevalent on the Web and outnumber (12,917 vs. 10,195) conventional SERP posts (P_1A_1). Also, in general, MCs produced more URIs than conventional SERP posts (P_1A_1). Additionally, MCs produced more non-HTML URIs than P_1A_1 across all topics. In fact, the total number of P_1A_1 non-HTML URIs were between 19% to 44% the size of MCs. These findings are potentially consequential for curators interested in enriching their collections with non-HTML resources.

From Table 15, for all topics in the Reddit SERPs except (Reddit-New), P_nA_n mostly produced the largest count of URIs (41,160), next to P_1A_1 (51% P_nA_n), next to P_nA_1 (3% P_nA_n): $P_nA_n > P_1A_1 > P_nA_1$. The relatively low number of Reddit P_nA_1 posts and URIs shows that it is a rare phenomenon for a Reddit user to reply to his/her initial post especially since Reddit does not impose any size restriction on the length of posts. For the Reddit-New SERP, P_1A_1 had more URIs (8,056) than P_nA_n (80% P_1A_1): $P_1A_1 > P_nA_n > P_nA_1$. This is likely due to the fact that the *New* SERP is constantly supplied with

new posts, so conversation thread among multiple users ($\mathbf{P}_n\mathbf{A}_n$ posts) are displaced as news posts ($\mathbf{P}_n\mathbf{A}_1$) arrive, since the *New* vertical is in “newest first” order. Consequently, before $\mathbf{P}_n\mathbf{A}_n$ sufficiently grow, they are pushed down (rank demotion) by newer $\mathbf{P}_1\mathbf{A}_1$ posts, and do not get sufficient exposure, leading to fewer replies which leads to a reduced $\mathbf{P}_n\mathbf{A}_n$ size.

The results show a high degree of inter/extra-user engagement on Twitter, and thus for Post and URI Counts (Table 15), $\mathbf{P}_n\mathbf{A}_n > \mathbf{P}_n\mathbf{A}_1 > \mathbf{P}_1\mathbf{A}_1$. In contrast, Scoop.it showed less user engagement, and thus: $\mathbf{P}_1\mathbf{A}_1 > \mathbf{P}_n\mathbf{A}_1$.

5.4.2 PROBABILITY DISTRIBUTION OF POSTS WITH LINKS

From Table 16, unsurprisingly, the probability of the event that a social media post with a URI of a given post class ($\mathbf{P}_1\mathbf{A}_1 - \mathbf{P}_n\mathbf{A}_n$) had more than one HTML URI ($k > 1$) seemed to correlate with whether the social media platform restricts the size of posts. For example, due to the character limit imposed on tweets, the probability of the event that a tweet with a URI has only 1 HTML URI is 0.98 ($P(p_{P_1A_1}^{Twitter} = 1) = 0.98$). On the other hand, single tweets with 3+ HTML URIs are rare. We observed three tweets with 3 or 4 HTML URIs (out of 3,501 tweets).

5.4.3 PRECISION OF POST CLASS URIS

Table 17 shows the conditional probability of the event that the URIs contained in a post of a given post class are relevant, given that the post has a specified count of URIs (k). Across almost all k per post class, we see that the seeds generated from $\mathbf{P}_1\mathbf{A}_1$ posts had a higher probability (maximum: 1.0, median: 0.63, minimum: 0.33) of being relevant than **MC** (0.61, 0.5, 0.0). For example, for Reddit when $k = 2$, $\mathbf{P}_1\mathbf{A}_1 - 0.80$, while **MC** - 0.59. This shows that $\mathbf{P}_1\mathbf{A}_1$ posts benefit from SERP filters; $\mathbf{P}_1\mathbf{A}_1$ posts are posts directly returned by SERPs and their text often matches a subset of the query. This indicates that a match between a query and a post text lends some relevance to the URI extracted from the post. However, given the fact that **MCs** do not all benefit from SERP filters since the vast majority of **MCs** are not extracted directly from the SERP, but from the reply or comment threads, the 0.5 median precision value indicates that comments and replies possess quality URIs.

In general, $\mathbf{P}_1\mathbf{A}_1$ post URIs (all URIs, HTML, and non-HTML) had the highest average precision compared to $\mathbf{P}_n\mathbf{A}_1$ and $\mathbf{P}_n\mathbf{A}_n$ for Reddit, Scoop.it, and Twitter posts extracted with text queries. For tweets extracted with hashtags, $\mathbf{P}_n\mathbf{A}_n$ posts had the highest average precision compared to $\mathbf{P}_n\mathbf{A}_1$ and $\mathbf{P}_1\mathbf{A}_1$.

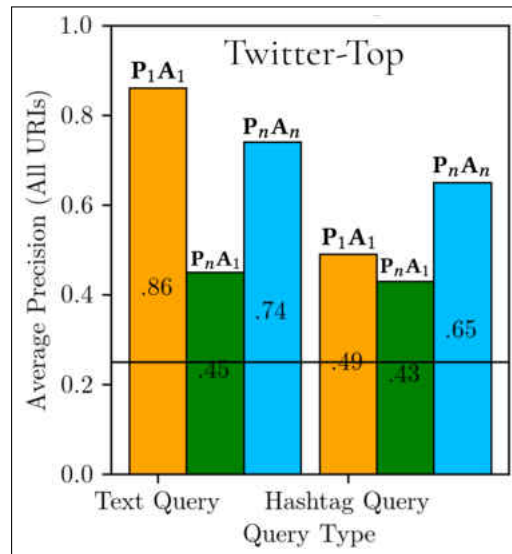


Fig. 33: Ebola Virus Outbreak Precision Distribution: P_1A_1 seeds produced webpages with a higher precision than P_nA_n for text but not hashtag queries. The black line marks the relevance threshold.

For Reddit, $P_1A_1 > P_nA_1 > P_nA_n$: across all topics, P_1A_1 posts had the highest average precision (all URIs) 80% of the time than P_nA_1 and P_nA_n . The Maximum, Median, and Minimum (MMM) average precision values were 0.88, 0.59, and 0.15, respectively. Next, P_nA_1 posts had a higher average precision than P_nA_n 70% of the time. The MMM of P_nA_n was 0.88, 0.50, and 0.00, respectively, and for P_nA_n is was 0.70, 0.42, and 0.07, respectively.

For tweets exposed with text queries, $P_1A_1 > P_nA_n > P_nA_1$: P_1A_1 (0.91, 0.66, 0.45) had the highest average precision 90% of the time than P_nA_n and P_nA_1 . P_nA_n (0.74, 0.46, 0.28) had a higher average precision 70% of the time than P_nA_1 (0.58, 0.39, 0.35).

For tweets exposed with hashtags, $P_nA_n > P_nA_1 > P_1A_1$: P_nA_n (0.65, 0.29, 0.27) posts had the highest average precision 60% of the time than P_nA_1 and P_1A_1 . P_nA_1 (0.45, 0.39, 0.21) posts had a higher average precision than P_1A_1 (0.50, 0.26, 0.11) 70% of the time. For example, from Figure 33, the average precision for P_1A_1 URIs in the Twitter-Top vertical for the *Ebola virus outbreak* topic was 0.86 (P_nA_n - 0.74) for posts extracted with the text query “ebola virus outbreak.” However, P_nA_n outperformed P_1A_1 (0.65 vs. 0.49) when the query used to extract posts was the hashtag “#ebolavirus.”

For Scoop.it, $P_1A_1 > P_nA_1$: P_1A_1 (0.87, 0.78, 0.55) posts had a higher average precision than P_nA_1 (0.80, 0.55, 0.27) 100% of the time. Similar to Twitter P_1A_1 , Scoop.it P_1A_1

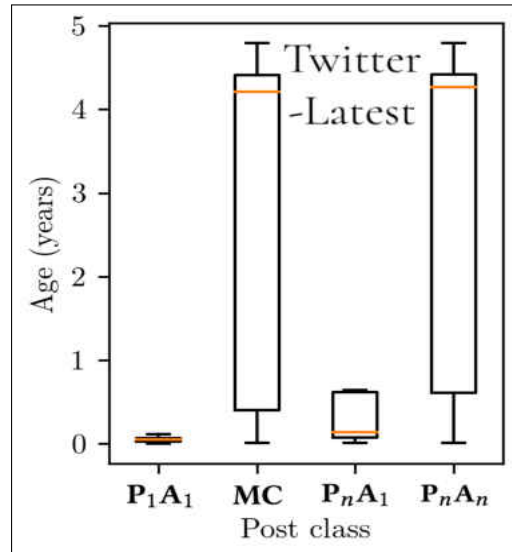


Fig. 34: Ebola Virus Outbreak Age Distribution: MCs produced older webpages in the Twitter-Latest vertical for the older topics.

are derived directly from the SERP, and thus benefit from SERP filtering. P_nA_1 do not benefit from SERP filtering since they are not extracted directly from the SERP.

5.4.4 AGE DISTRIBUTION OF RELEVANT WEBPAGES

We compared the ages of P_1A_1 and MC post class URIs, by focusing on the older topics (*Ebola virus outbreak* and *Flint Water Crisis*) for social media that supports P_1A_1 , P_nA_1 , and P_nA_n - Reddit and Twitter. MC posts consistently produce older webpages in the Twitter-Latest vertical. A possible explanation for this is that P_1A_1 tweets (extracted directly from the Twitter-Latest SERP) are highly likely to be new tweets if the topic is ongoing. Even though new tweets can include URIs of old stories, for ongoing news stories such as those we considered, new tweets are likely to include the URIs of the latest developments. We observed that the Twitter-Latest P_1A_1 tweets were created within days from the query issue dates, and thus were more likely to produce new URIs for both topics. In contrast, MCs are extracted from conversations that can mix new and old tweets; a new tweet can reply to an old tweet that contains old URIs. Therefore, Twitter-Latest MCs produced a mix of tweets created within days and years from the query issue dates.

For the Reddit-Top/Relevance/Comments SERPs for *Ebola virus outbreak*, MCs and P_1A_1 produced older webpages with similar distributions. For example, for *Ebola virus*

outbreak both post classes had a median webpage age of 4.3 years.

As expected, the Reddit-New, for both topics, **MCs** and $\mathbf{P}_1\mathbf{A}_1$ produced the newest webpages compared to other Reddit SERPs with median age < 1 year.

$\mathbf{P}_1\mathbf{A}_1$ and **MC** posts from Twitter-Top produced webpages with similar age distributions. For example, for *Flint Water Crisis* both post classes had a median webpage age < 5 months. In contrast, in the Twitter-Latest vertical, for both topics **MCs** produced older webpages than $\mathbf{P}_1\mathbf{A}_1$. For example, **MCs** for *Ebola virus outbreak* produced older webpages (median: 4.2 years) than those from $\mathbf{P}_1\mathbf{A}_1$ (19 days) (Figure 34).

5.4.5 DISTRIBUTION OF HOSTNAME DIVERSITY

For Reddit, the $\mathbf{P}_n\mathbf{A}_1$ posts produced the highest hostname diversity. For Twitter, $\mathbf{P}_1\mathbf{A}_1$ posts produced the highest hostname diversity.

For Reddit, $\mathbf{P}_n\mathbf{A}_1 > \mathbf{P}_1\mathbf{A}_1 > \mathbf{P}_n\mathbf{A}_n$: across all topics, $\mathbf{P}_n\mathbf{A}_1$ posts had the highest hostname diversity (HTML URIs) 95% of the time than $\mathbf{P}_1\mathbf{A}_1$ and $\mathbf{P}_n\mathbf{A}_n$. The Maximum, Median, and Minimum (MMM) hostname diversity values were 1.0, 0.55, and 0.0, respectively. Next, $\mathbf{P}_1\mathbf{A}_1$ posts had more diverse hostnames than $\mathbf{P}_n\mathbf{A}_n$ 61% of the time, MMM - (0.6, 0.33, 0.11), for $\mathbf{P}_n\mathbf{A}_n$ - (0.55, 0.28, 0.1).

For Twitter, $\mathbf{P}_1\mathbf{A}_1 > \mathbf{P}_n\mathbf{A}_n > \mathbf{P}_n\mathbf{A}_1$: $\mathbf{P}_1\mathbf{A}_1$ (0.70, 0.60, 0.43) produced more diverse hostnames 74% of the time than $\mathbf{P}_n\mathbf{A}_n$ and $\mathbf{P}_n\mathbf{A}_1$. Similarly, $\mathbf{P}_n\mathbf{A}_n$ (0.61, 0.45, 0.39) produced more diverse hostnames 79% of the time than $\mathbf{P}_n\mathbf{A}_1$ (0.74, 0.37, 0.31). Scoop.it did not produce enough URIs for two topics, as a result had fewer $\mathbf{P}_n\mathbf{A}_1$ to derive a fair comparison with $\mathbf{P}_1\mathbf{A}_1$.

Reddit and Twitter had $\mathbf{P}_1\mathbf{A}_1 > \mathbf{P}_n\mathbf{A}_n$ in common. This is not unexpected; hostname diversity rewards unique hosts, and given that the $\mathbf{P}_1\mathbf{A}_1$ collection is smaller than $\mathbf{P}_n\mathbf{A}_n$, it is more likely for $\mathbf{P}_1\mathbf{A}_1$ to fill in the hostname slots with additional different hosts than $\mathbf{P}_n\mathbf{A}_n$. However, for Twitter $\mathbf{P}_n\mathbf{A}_1$ had the lowest diversity unlike Reddit for the following reasons. First, $\mathbf{P}_n\mathbf{A}_1$ is the set of all threads authored by the same user. These threads on Twitter, especially those from news (e.g., @nytimes, @vice) and non-news organizations (e.g., @splcenter, @TurnoutPAC) tend to link to webpages within their websites, leading to a lower hostname diversity. This phenomenon was most prominent in the 2018 *World cup* and *Midterm elections* topics.

5.4.6 OVERLAP: GOOGLE COLLECTIONS VS. POST CLASSES

All post classes showed small amount of overlap with the collections of URIs returned

TABLE 18: Summary recommendations of the **Source** to prioritize when generating seeds from social media based on the **Attribute Prioritized**, **Query Type**, and **Vertical**.

Attribute Prioritized				Query Type	Vertical	Source
Quantity	Quality	Older Seeds	Hostname diversity			
Yes	No	N/A	N/A	Text	Top	MC
No	Yes	N/A	N/A	Text	Top	P₁A₁
Yes	No	N/A	N/A	Hashtag	Top	MC
No	Yes	N/A	N/A	Hashtag	Top	MC
N/A	N/A	Yes	N/A	Text/Hashtag	Latest	MC
N/A	N/A	N/A	Yes (For Twitter)	Text/Hashtag	Top/Latest	P₁A₁

from the first 10 pages of Google for the respective dates the post class URIs were extracted. This highlights the fluidity of the Google SERP. Thus, URIs extracted from **MC** and **P₁A₁** collections are not easily discoverable.

Reddit **P₁A₁** and **MC** posts had overlap < 0.1 85% of the time. Their MMM overlap were: 0.13, 0.04, and 0.1, respectively. Twitter **P₁A₁** posts had overlap $(0.09, 0.02, 0.0) < 0.1$ 100% the time. Similarly, Twitter **MC** posts had overlap $(0.13, 0.04, 0.0) < 0.1$ 80% of the time.

5.5 GENERATING SEEDS FROM SOCIAL MEDIA, A RECOMMENDATION

Considering the results presented in Chapter 5.4, it is clear that collections generated from social media SERPs (**P₁A₁**) are different from collections generated from Micro-collections (**MCs**), and both post classes yield seeds not easily discoverable by scraping Google. Consider the following recommendations, described next, and summarized in Table 18, about the sources (**MC** or **P₁A₁**) to focus on when generating seeds, based on the attribute (quantity, quality, age, and domain diversity) the user prioritizes.

MCs are more prevalent and produce more seeds than **P₁A₁**. This means seed generation that prioritizes quantity would benefit from extracting seeds from **MCs**. **P₁A₁** produced higher quality URIs for all social media SERP combinations except with seeds generated with hashtags. The poorer precision performance of hashtag queries compared to text queries shows that hashtags can be used as a vehicle for spreading non-relevant content, especially

when the hashtag is popular. However, when users reply to a tweet that contains a link and a hashtag (the composition of $\mathbf{P}_n\mathbf{A}_n$ set), it is likely they are responding to a relevant tweet. Replies may serve as a quality check. Therefore, **MCs** produced more relevant URIs when hashtags were used to surface tweets. Consequently, seed generation that prioritizes quality would benefit from extracting seeds from $\mathbf{P}_1\mathbf{A}_1$, but for Twitter, if hashtags are used, **MCs** should be considered first.

MCs consistently produced older webpages than $\mathbf{P}_1\mathbf{A}_1$ posts for the Twitter-Latest vertical because **MCs** included older tweets. Consequently, if seed generation from the Twitter-Latest vertical intends to extract older stories, **MCs** should be prioritized. Finally, we showed that $\mathbf{P}_1\mathbf{A}_1$ produced more diverse hostnames than **MCs** for Twitter unlike Reddit. Therefore, seed generation that intends to include different hosts should consider $\mathbf{P}_1\mathbf{A}_1$, instead of Twitter $\mathbf{P}_n\mathbf{A}_1$, since it showed a low level of hostname diversity due to resampling of the same domains, which is a common practice especially among news organizations.

5.6 CHAPTER SUMMARY

This chapter first introduced a vocabulary (post class) for labeling social media posts regardless of platform, based on the number of posts and authors. Next, we introduced Micro-collections on social media as social media posts that exhibit collection building attributes such as selection and filtering. We compared seeds generated using conventional methods such as scraping social media SERPs with Micro-collections ($\mathbf{P}_1\mathbf{A}_1$ vs **MC** post class) and showed that both methods lead to different kinds of seeds. For example, Micro-collections yield more URIs than conventional scraped SERP seeds, but with less precision. However, for hashtag queries, Micro-collections produced URIs with a higher precision than scraped SERP seeds, showing that hashtags may be used as a vehicle for spreading non-relevant content, and replies may be used as a quality check.

CHAPTER 6

COMPARING COLLECTIONS OF SEEDS

Given two collections of seeds A and B , each containing news reports about the *2014 Ebola virus outbreak*, consider that collection A has 50 URIs and was manually collected by a pathologist on August 4, 2014. Collection B contains 50 URIs extracted by issuing the “ebola virus” query to Twitter on the same day and extracting the first 50 relevant links embedded in the tweets. How do we characterize these collections? Can we compare them? In this chapter, we take a first step toward answering these critical questions and make contributions to address the third research question (Chapter 1.4) of this effort, repeated here for convenience:

- **RESEARCH QUESTION 3:** How do we evaluate automatically created collections with those generated by human experts in Archive-It?

It is not sufficient to generate seeds, it is also necessary to determine the nature of, or to characterize the seeds collected, and possibly compare them to other seeds collected around the same topic. This chapter presents our study [136] to investigate how to characterize and compare collections. Comparing collections is challenging because it requires comparing collections that may cater to different needs. It is also challenging to compare collections since there are many possible measures to use as a baseline for collection comparison: how does one narrow down this list to metrics that reflect if two collections are similar or dissimilar? We addressed these challenges in two main steps. First, we explored the state of the art in collection comparison and defined a suite of seven measures, called Collection Characterizing Suite (CCS), discussed in Chapter 6.1 to describe the individual collections. Second, to compare collections (Chapter 6.2), we calculated the distances between the CCS vectors of collections. We applied our method to check if collections generated automatically and semi-automatically from social media sources such as Storify, Reddit, Twitter, and Wikipedia are similar to Archive-It human-generated collections. Our results showed that social media sources such as Reddit, Storify, Twitter, and Wikipedia produce collections that are similar to Archive-It collections. Consequently, curators may consider extracting URIs from these sources in order to begin or augment collections about various news topics.

6.1 COLLECTION CHARACTERIZING SUITE (CCS)

The CCS provides a means of characterizing individual collections and comparing multiple collections across seven dimensions (Chapter 6.1.1 – Chapter 6.1.7):

1. Distribution of topics
2. Distribution of sources
3. Content diversity
4. Temporal distribution
5. Source diversity
6. Collection exposure
7. Target audience

The various metrics that make up the CCS can be instantiated in different ways - it is a template. Consequently, the main criteria considered for instantiating the various metric was generality.

6.1.1 DISTRIBUTION OF TOPICS

A “topic” is informally defined as a group of words which frequently occur together. It provides a means to summarize collections and gives us some notion of what the collection is about. It is impractical to manually inspect all the webpages, especially for large collections, in order to discern aboutness, therefore, we need this measure to summarize collections. The distribution of topics is a ranked list of topics in a collection with the most frequent topics (most important summaries) at the top and the least frequent topics (least important summaries) at the bottom. A probabilistic language model assigns probabilities to a sequence of words that make up a topic. One goal of a language model is the assignment of high probabilities to frequent topics (or sentences) in a collection. Similarly, we adopted a variant of the n-gram language model. Since collections are organized around specific topics, webpages in the collection include these topics frequently in their vocabulary. For example, we would expect a collection about *sports* events to possess sports vocabulary, e.g., *football* and *basketball*. Inspired by this characteristic of collections, we developed a method to derive the topical distribution of a collection by finding the n-grams in the collections

Algorithm 2 : Generate a distribution of n-grams (topics)

Input: A collection C of webpages ($|C| = N$), integers $n > 0$, & $m > 0$.

Output: A ranked list of m n -grams (topics); the n -grams with the highest frequencies at the top of the list.

Function GenTopicDist(C, n, m)

0. Represent each document $d_i \in C$ as a n -gram document

1. Create a vocabulary vector $V \in \mathbb{Z}^{1 \times p}$, each entry v_i in V represents a unique n -gram from C (with p unique n -grams).

2. Create a binary document term matrix $\mathbf{M} \in \mathbb{Z}^{N \times p}$. Each row in \mathbf{M} represents a document $d_i \in C$, and each column has 1 if $v_i \in d_i$, and 0 otherwise.

3. Create a ranked list L . Populate L ($|L| \leq m$) with n -grams (v_i) with the highest frequencies of occurrence in \mathbf{M} ($\max_{v_i \in V} \sum_{j=1}^N m_{j,i}$).

Populate L with v_i in decreasing order of their frequencies.

return L

EndFunction

with the highest frequency of occurrence in the collection. The method is described by Algorithm 2 and sample outputs are given in Table 19. Algorithm 2 leads to the possibility of splitting multi-word proper nouns n-grams. For example, given an *Ebola virus* collection, if we choose $n = 2$ to generate bigram topic distributions, it could result in a ranked list that includes “centers disease” and “disease control”. It is clear that both terms are part of the multi-word proper noun (trigram) “centers disease control” (stopwords are removed). To solve this problem, we replace multiple lower-order (e.g., bigram) n-grams with their superset higher-order (e.g., trigram) n-grams. Algorithm 2 has been further developed, optimized, and reimplemented as a Python application called *sumgram* [160].

6.1.2 DISTRIBUTION OF SOURCES

Given a collection of webpages, the distribution of sources is a statistical summary of the various sources sampled in order to build the collection. For example, the NLM Archive-It *Ebola virus* collection [3] consists of 18 (12.5%) webpages from `blogs.plos.org`, 14 (9.7%) from `cdc.gov`, and 11 (7.6%) from `twitter.com`. We may conclude that these are the three

TABLE 19: Distribution of Top Five Topics for Two Archive-It Collections.

2016 Pulse Nightclub Shooting	Hurricane Harvey
“pulse nightclub orlando florida”	“hurricane harvey photo”
“new york”	“27 2017 houston”
“en la comunidad”	“27 2017 photo”
“mass shooting”	“tropical storm harvey photo”
“omar mateen”	“corpus christi”

most influential sources in the collection.

The distribution of sources is instantiated with a simple enumeration of the frequencies of the various hosts that make up a collection. In order to make the description more compact, we chose to report the top 10 hosts that make up a collection, and what proportion of the collection the top 10 hosts account for. For example, the top 10 hosts in the NLM Archive-It *Ebola Virus* collection make up 50% of the collection.

6.1.3 CONTENT DIVERSITY

Given a collection of webpages, the content diversity is defined as the degree of variety of the content of the webpages in the collection. For example, if we sample a collection about a *shooting* event one hour after the event, we should expect a low degree of variety (high degree of similarity) in the webpages. Most of them are expected to report the location of the shooting, the casualty count, possible identity of the perpetrators, etc. However, one year after the event, we may see more diverse content, perhaps discussing the shooting in context to other shootings. The diversity of the content of such events increases with time.

A diversity score of 0.0 means no diversity (duplicate web documents), and a diversity score of 1.0 means maximum diversity (mutually orthogonal vocabulary of documents). Content similarity, the opposite of content diversity, is a useful metric that has been applied to quantify the similarity of news stories which often occurs after a major news event [161].

The input to calculate content diversity for an arbitrary collection is a similarity matrix D . The similarity matrix consists of the pairwise similarity of the web documents in the collection. We propose two ways of calculating the similarity between a pair of webpages corresponding with the two different ways of representing a collection. First, a collection may be represented as a *Document-Term matrix*: each row represents a document (webpage) and

each column represents the TF or TFIDF value of a unigram in the collection vocabulary. In this representation, the similarity between a pair of documents is the cosine similarity measure. Second, a collection may be represented as a *List of Entity sets*: each document is represented as a set of entities of proper nouns (**people**, **location**, **organization**, **time**, **date**, **money**, **percent**, and **misc**). The entities were extracted using the Stanford Named Entity Recognition System [162]. In this representation, we defined a new similarity measure - weighted Jaccard-Overlap similarity (Equation 7) to calculate the similarity between a pair of web documents, with a Jaccard weight ($\alpha \in [0, 1]$) of 0.4.

The weighted Jaccard-Overlap similarity $sim(A, B)$ between a pair of documents sets A and B is given by Equation 7, where β is the coefficient of similarity, defining the threshold two documents must reach to be considered similar. This threshold was empirically derived from a gold-standard dataset and set to 0.27. Specifically, to set the threshold, we manually created 20 collections of news stories from multiple topics (*politics*, *entertainment*, etc.). Each collection consisted of multiple URIs representing the same news story reported by multiple news organizations. Next, for different thresholds, we clustered the collections of news stories with Equation 7 and identified the threshold (0.27) that minimized the clustering error.

$$sim(A, B) = \begin{cases} 1 & , \text{ if } \alpha \cdot J(A, B) + (1 - \alpha) \cdot O(A, B) \geq \beta \\ 0 & , \text{ otherwise} \end{cases} \quad (7)$$

$J(A, B)$ is the Jaccard index of both documents, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and $O(A, B)$ is the Overlap coefficient of both documents, $O(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$.

Let a similarity matrix of n webpages in a collection be represented by $\mathbf{D} \in \mathbb{R}^{n \times n}$, and an *all-ones matrix* $\mathbf{O} \in \mathbb{R}^{n \times n}$. Given a square matrix, $\mathbf{N} \in \mathbb{R}^{n \times n}$, with zeros on the main diagonal and ones everywhere else. For example, if $\mathbf{N} \in \mathbb{R}^{3 \times 3}$,

$$\mathbf{N} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ the content diversity score } d_c = 1 - \frac{\|\mathbf{ND}\|_F}{\|\mathbf{NO}\|_F}$$

where $\|\mathbf{A}\|_F$ is the Frobenius norm: $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2}$. Web documents consist of topics (groups of words that frequently occur together). This means multiple words that belong to the same topic tend to co-occur. We may not always consider our collection diverse by the mere presence of different words, especially if these words belong to the same topic. Instead, we may consider our collection diverse if it consists of different topics.

Consequently, if we consider unigrams, we would reward diversity to different terms which occur together, even though they may belong to the same topic, i.e., no new information. The *Document-Term matrix* representation rewards diversity at the term level, while the *List of Entity sets* representation rewards diversity at the topic level.

6.1.4 TEMPORAL DISTRIBUTION

The *publication temporal distribution* is an aggregation of publication dates that are used to timestamp webpages. The *content temporal distribution* is the collection of time references associated with events being discussed on webpages. The time information may be absolute, (e.g., “On Friday, Nov 17, 2017...”) or relative (e.g., “Next month is...”). We normalize relative time information (e.g., if the reference date is “2017-11-17” we represent “next month” as “2017-12-17”). Temporal distributions enable the calculation of the collection age. The ages of webpages may be calculated with respect to the creation date of the collection to indicate how long webpages existed prior to being collected. A short duration between the publication date of webpages and the creation date of the collection may indicate that the curator intended to collect webpages following a recent event. Alternatively, the ages of documents may be calculated with respect to the current date to determine absolute ages of webpages.

The publication dates of webpages may provide useful information about the kinds of events discussed in the document. For example, stories concerning airport security before the September 11, 2001 terrorist attacks are not expected to discuss the TSA (Transportation Security Administration), because the TSA was founded on November 19, 2001. The publication date alone may not be sufficient to give us a full picture of the kinds of events discussed in a document, since documents often discuss events and include the dates of these events in their content. This may be relative, e.g., “last year” or absolute “on Jan 3rd, 2017.” Therefore, we also have to pay attention to these dates.

We extract the publication dates of the documents in a collection to form the publication date distribution through the use of CarbonDate [159] which estimates the creation date of webpages based on information polled from multiple sources such as the document timestamps, web archives, Twitter, backlinks, etc. We extracted the content dates with the aid of SUTime [163].

6.1.5 SOURCE DIVERSITY

Similar to content diversity, the source diversity metric tells us whether a collection

samples a single source, a handful of sources, or many sources. The URI source diversity metric [164], $d_{URI} \in [0, 1]$ tells us the rate of unique URIs; $d_{URI} = 0.0$ means the collection only has one distinct URI (duplicate webpages). On the other hand, if $d_{URI} = 1.0$, it means the collection is made up of unique URIs. We also explore source diversity at the domain (d_{domain}) and hostname ($d_{hostname}$) policies.

We deduplicated URIs in collections by trimming all parameters from the URIs as suggested by Brunelle et al. [144] before calculating source diversity. Given a policy set $P = \{URI, Domain, Hostname\}$ for a collection C and the count of unique URIs in the collection U , the source diversity of a given policy d_p is given by Equation 8.

$$d_{p \in P} = \frac{U}{|C|}; d_p \in [\frac{1}{|C|}, 1] \quad (8)$$

The normalized source diversity of a given policy d'_p is given by Equation 9.

$$d'_{p \in P} = \frac{d_p - \frac{1}{|C|}}{1 - \frac{1}{|C|}} = \frac{U - 1}{|C| - 1}; d'_p \in [0, 1] \quad (9)$$

The social media diversity metric or social media rate quantifies the proportion of webpages in a collection that are from social media sites. We created a predefined list of social media domains: `twitter.com`, `facebook.com`, `youtube.com`, `instagram.com`, and `tumblr.com`. Given k URIs from social media domains in a collection C , the social media rate is $\frac{k}{|C|}$. For example, a collection composed of 3 URIs from Twitter, 2 from Facebook, and 5 from CNN, has a social media rate of $\frac{3+2}{10} = 0.5$.

6.1.6 COLLECTION EXPOSURE

If a webpage is “popular” (used widely), this means there is some need that the document fulfills for a wide audience. We approximate popularity with the collection exposure metrics, *archival rate* and *tweet index rate*. In our previous work [142], we showed that collections of local news from local news organization, such as the *Caloosa Belle newspaper* (LaBelle, Florida USA), are less exposed, thus less popular than collections of news sources from mainstream news organizations, such as *CNN* and *The Washington Post*.

The archival rate of a collection C is the fraction of C that is archived. For example, if we found 10 archived stories from C (where $|C| = 50$), the archival rate of C is $\frac{10}{50} = 0.2$. Note that when comparing the archival rates of two collections, it is important to consider how old both collections are. For example, collection A might have a much larger archival rate than collection B only because A has much older documents than B , and as a result had the greater opportunity to be archived.

Popular (widely used) URIs are more likely to be archived than less popular URIs [165]. This means we could use the archival state of a URI to infer its popularity. This method will not be valid if every URI is archived (e.g., Archive-It seeds). If this were the case (all URIs archived), the magnitude of archived copies of a URI may indicate its popularity. The archive state of a webpage can be measured using Memgator [166].

Similar to the archival rate, the tweet index rate of a collection C is the fraction of C found embedded in tweets. For example, if we found 40 URIs from C (where $|C| = 50$) embedded in tweets, the tweet index rate of C is $\frac{40}{50} = 0.8$. Also similar to archival rate, when comparing the tweet index rates of two collections, it is important to consider how old both collections are. For example, collection A might have a much larger tweet index rate than collection B only because A includes webpages that are much older than B , and as a result, had a greater opportunity to be tweeted. The tweet index state of a webpage is set by searching Twitter for a tweet that embeds the page URI [167].

Similar to the archival rate, popular URIs are more likely to be shared on social media sites like Twitter than less popular URIs. Consequently, the tweet index state (in tweet or not) of a webpage may indicate the popularity or exposure of the webpage. We may also be able to infer the popularity of a URI in a tweet by taking into account how often it is shared on Twitter. The tweet index rate is often a useful alternative to the archival rate when the collections to be compared have the same archival rate. For example, Archive-It seeds have a 100% archival rate. Likewise the archival rate provides an alternative when comparing collections with the same tweet index rates, for example, collections generated from Twitter have 100% tweet index rates.

6.1.7 TARGET AUDIENCE

The target audience estimates the target users of the collection. This is not easy to achieve. Our premise is that the readability level of the documents in the collection is a reflection of the target audience. For example, if the reading level of a collection is at the 10th grade level, we conclude that the target audience starts from high school young adults and above. However, if the reading level is at the graduate level (16th grade) level, we may conclude the target audience might be professionals in a subject area.

The target audience of a collection provides important contextual information that may give insight about the composition of the collection, and may reflect the intent of the collection builder, such information is not often readily available.

We instantiate the target audience metric with readability measures. Readability measures estimate the reading level of documents through procedures that include counting syllables, words, and sentences. We employed three widely used readability measures that output grade levels: the Flesch-Kincaid Grade level [168], Coleman Liau index [169], and the Automated Readability index [170]. For a single document, the readability score is the average score from the three readability measures (normalized between 0 and 1). The higher the readability score, the higher the grade level.

6.2 COLLECTION CHARACTERIZATION AND COMPARISON

In order to characterize a single collection with the CCS, we simply instantiate the metrics that make up the suite. The state of the metrics collectively form a characterization for the collection. For example, Table 20 describes two collections. The first, the NLM Archive-It *Ebola Virus* collection, is an archived collection built manually by an archivist at the NLM in October 2014. The second, the Reddit *Ebola Virus* collection, we built by issuing the query “ebola virus” to Reddit from 2017-07-25 to 2017-08-23 and extracting links from the Reddit SERPs and their respective comments. Let us consider both collections to see how the CCS describes collections.

The top five topics from the NLM Archive-It collection show that the collection addresses issues arising from the Ebola virus outbreak in West Africa (Table 20a, topic 1) and that the main countries affected were Guinea, Liberia, and Sierra Leone (Table 20a, topic 2). Also two major players involved with the outbreak were public health workers and the Centers for Disease Control and Prevention (Table 20a, topic 4 & 5). The Reddit collection also mirrors this sentiment. Both collections are similarly characterized by the fraction of the collections the top 10 hosts make (Table 20b, Distribution of sources). Similarly, both collections target a similar audience (Table 20b, Target audience) since they have the same median normalized grade level of 0.57 (11th grade).

Table 20b shows that the Reddit collection produced a higher content diversity for both collection representations (*Document-Term matrix* and *List of Entity sets*). The NLM Archive-It collection produced much newer web documents with a median publication age of 36 days, compared to the Reddit collection of 3.9 years. This suggests that the NLM Archive-It collection was created a few months after the Ebola event unfolded. Additionally, the Reddit collection sampled from more hosts (hostname source diversity - 0.53) and had more social media URIs (social media rate - 0.12) compared to the NLM Archive-It collection (hostname source diversity - 0.34, social media rate - 0.07). The NLM Archive-It collection

NLM (occurrence rate)	Reddit (occurrence rate)
“ebola outbreak west africa” (0.34)	“infected ebola virus disease” (0.25)
“guinea liberia sierra leone” (0.31)	“west africa” (0.21)
“cases ebola virus disease” (0.30)	“public health workers” (0.15)
“public health workers” (0.27)	“sierra leone” (0.15)
“centers disease control prevention” (0.15)	“united states” (0.14)

(a) Distribution of top five topics for NLM Archive-It and Reddit *Ebola virus* collections showing a similar topic distribution.


CCS Metric	NLM Ebola Characterization	Reddit Ebola Characterization
Distribution of sources	Top 10 hosts fraction of collection: 50%	Top 10 hosts fraction of collection: 46%
Content diversity (Doc-Term matrix / Entity set)	(0.80 / 0.65)	(0.89 / 0.85)
Publication temporal dist. (Median age, where age: Creation date - Pub. date)	36 days	1,450 days (3.9 years)
Content temporal dist. (Median age)	1,144 days (3.1 years)	2,104 (5.8 years)
Source diversity (URI/ Hostname / Social media)	(1.0 / 0.34 / 0.07)	(0.98 / 0.53 / 0.12)
Collection exposure (Archival rate/ Tweet index rate)	(1.00 / 0.72)	(0.78 / 0.40)
Target audience (readability, Q1 / Median / Q3)	(0 / 0.57 / 1)	(0.14 / 0.57 / 0.85)

(b) CCS characterizations of NLM and Reddit *Ebola virus* collections

TABLE 20: Characterization of two collections Archive-It (144 URIs) and Reddit (150 URIs) *Ebola virus* collections. Each characterization describes the individual collection, juxtaposing multiple characterizations enables collection comparison.

indicated a higher exposure than the Reddit collection, with a higher archival rate of 1.0, compared to the 0.78 archival rate of the Reddit collection. The high archival rate of the Archive-It collection is no surprise because it is a collection of seeds; the seeds are meant to be crawled and archived. The NLM Archive-It collection also showed a higher tweet index rate (0.72) than the Reddit collection (0.40).

6.3 EVALUATION

To assess if we could bootstrap archived collections from social media, we measured the distances between archived collections from Archive-It () and collections generated from social media sources: Storify () , Reddit () , Twitter Moments () , Twitter SERP () , and Wikipedia () . The rationale for this is, if collections created by extracting URIs from social media collections are similar (low distance) to expert-created collections on Archive-It, then we may start or augment archived collections with seeds extracted from social media sources.



























We generated a dataset (Table 21) of 129 collections (2,765 URIs) from three topics: “Ebola Virus,” “Hurricane Harvey,” and “2016 Pulse Nightclub Shooting,” and 10 collections (500 URIs) for random (multiple topics) news stories from the UCI news aggregator dataset [171]. Random collections () were included to assess if the CCS resulted in clusters of collections of common topics even in the presence of noise. We do not expect collections of random news stories to be more similar to archived collections than social media collections. Additionally, we included baseline collections generated by extracting URIs from Google () . We believe most users primarily use Google to discover candidate URIs for their collections, so we included Google collections in order to quantify how these compare with social media and archived collections. Our previous work [24] showed that such collections change with time since search engines are biased to produce the latest documents. The evaluation dataset collections were represented as a vector of CCS values, and a distance was calculated between Archive-It collections (Table 21, IDs 1, 8, and 13) and every other collection irrespective of the topics. The Euclidean distance metric was used (as opposed to cosine) to compute distance because the magnitudes of the respective CCS values in the collection vectors are significant. We normalized (0 – 1) the Euclidean distances since all possible maximum and minimum CCS values are known. Additionally, the CCS metrics were assessed to identify the metrics which provided the most information in distinguishing the collections. This was done by calculating the spread of values (standard deviation) of the individual CCS metrics for the collections.

TABLE 21: The *CCS* Evaluation Dataset comprised of 129 collections from three Topics: “Ebola Virus,” “Hurricane Harvey,” and “2016 Pulse Nightclub shooting.” WSDL represents the collections generated by the authors.

ID	Topic (URI Count)	Source (Author)	Creation Date	Extraction note
1	Ebola Virus Outbreak (144)	 (NLM)	2014-10	Archive-It seeds
2	Ebola Virus Outbreak (669)	 (WSDL)	2017-11-29	100 sub-collections (IDs 0-99) of URIs from 100 Storify <i>stories</i>
3	Ebola Virus Outbreak (669)	 (WSDL)	2017-11-29	A Collection created by combining all links in Collection 2.
4	Ebola Virus Outbreak (155)	 (WSDL)	2017-07-25	URIs from references of <i>Ebola Virus</i> Wikipedia page
5	Ebola Virus Outbreak (153)	 (WSDL)	2017-07-25 - 2017-08-23	URIs from Reddit (& comments) search for query: “Ebola Virus”
6	Ebola Virus Outbreak (152)	 (WSDL)	2017-08-02 - 2017-11-28	URIs in tweets from Twitter search for query: “Ebola Virus”
7	Ebola Virus Outbreak (105)	 (WSDL)	2017-11-29	URIs from first 10 pages of Google, for query: “Ebola Virus”
8	Hurricane Harvey (44)	 (IA)	2017-08	Archive-It seeds
9	Hurricane Harvey (151)	 (WSDL)	2017-09-02	URIs from references of <i>Hurricane Harvey</i> Wikipedia page
10	Hurricane Harvey (14)	 (WSDL)	2017-12-08	2 sub-collections (IDs 0 – 1) of URIs from tweets in Twitter Moments
11	Hurricane Harvey (14)	 (WSDL)	2017-12-08	A collection created by combining all URIs in Collection 10
12	Hurricane Harvey (94)	 (WSDL)	2017-09-02- 2017-11-29	URIs from first page of Google, for query: “Hurricane Harvey”
13	2016 Pulse Night Club Shooting (151)	 (IA)	2016-06	Archive-It seeds
14	2016 Pulse Night Club Shooting (50)	 (WSDL)	2017-12-08	5 sub-collections (IDs 0-4) of URIs from tweets from Twitter Moments
15	2016 Pulse Night Club Shooting (50)	 (WSDL)	2017-12-08	A collections created by combining URIs in Collection 14
16	Random (500)	 (Lichman, M)	2014-03-10 - 2014-08-10	10 sub-collections (IDs 0-9) of URIs or random news stories
Total	2,765 URIs			






We generated a CCS matrix for the evaluation dataset collections. The rows of the CCS matrix represent the collections and the columns represented the CCS metric values. The first and second columns represent the content diversity values calculated with the *Document-Term matrix* and *List of Entity sets* collection representations, respectively. The third column represents the *URI source diversity*, fourth - *domain source diversity*, fifth - *hostname diversity*, sixth - *social media rate*, seventh - collection exposure *archival rate*, eighth - collection exposure *tweet index rate*, and ninth, the Jaccard similarity score of a given collection’s top 10 n-gram distribution of topics to the Archive-It collection. The last column of the CCS matrix represented the normalized median reading level of the collection. Chapter 6.1 outlines how to extract the CCS metrics of all the entries, except the Jaccard similarity of the n-gram distribution of topics for two collections. The idea for this method is to find how similar two collections are in terms of their respective n-gram distribution of topics, in other words, if the collections are about a similar set of topics. We focused on finding similar collections based on the content of the collection and not the sources they sample from or the time the collection was built. Consequently, we excluded the distribution of sources and temporal distributions from the CCS vector.

TABLE 22: List of collections most similar to three Archive-It collections and three random collections for the evaluation dataset topics.

Gold standard collections	Three most similar
Ebola Virus (ebo.)  <i>ebo.1</i> 0	 <i>ebo.5</i> 0.17  <i>ebo.2.49</i> 0.23  <i>ebo.2.57</i> 0.25
Hurricane Harvey (hur.)  <i>hur.8</i> 0	 <i>hur.12</i> 0.27  <i>hur.11</i> 0.32  <i>pul.15</i> 0.34
2016 Pulse night.. (pul.)  <i>pul.13</i> 0	 <i>pul.15</i> 0.24  <i>pul.14.2</i> 0.24  <i>pul.14.4</i> 0.31
Random news stories 0 (ran.)  <i>ran.16.0</i> 0	 <i>ran.16.8</i> 0.16  <i>ran.16.5</i> 0.19  <i>ran.16.4</i> 0.22
Random news stories 1 (ran.)  <i>ran.16.1</i> 0	 <i>ran.16.6</i> 0.19  <i>ran.16.3</i> 0.22  <i>ran.16.8</i> 0.22
Random news stories 2 (ran.)  <i>ran.16.2</i> 0	 <i>ran.16.3</i> 0.22  <i>ran.16.4</i> 0.22  <i>ran.16.5</i> 0.22

6.4 RESULTS

Each pictogram in Table 22 represents a collection expressed by an image of the collection source (Chapter 6.3). The pictogram superscript represents the collection topic abbreviation followed by the collection ID (Table 21). The sub-collection ID follows the collection ID for Storify and Twitter Moments sub-collections. The subscript represents the normalized Euclidean distance of the collection to the specified Archive-It collection. For example, for the *Ebola Virus* topic, the Reddit (*ebo.5*) collection had the closest distance (0.17) to the Archive-It (*ebo.1*) collection.

Table 22 shows that the CCS resulted in the clustering of collections of similar topics with a distance ranging from 0.17 to 0.34 across all topics. The Reddit collection ( ^{*ebo.5*}_{0.17}) was most similar to the Archive-It *Ebola Virus* collections ( ^{*ebo.1*}₀). Since we had more Storify collections in our dataset, the Storify collections have a higher opportunity of outperforming (lowest distance) other collections. In fact, the Storify *Ebola Virus* collection ( ^{*ebo.3*}_{0.27}) is 4.3 times the size of the Reddit collection, yet, the Reddit collection was most similar to the Archive-It collection. This suggests that the larger the collection may not always mean the better the collection. This result is potentially consequential: it suggests that we may consider Reddit as a collection source in the absence of Storify. The Google *Hurricane Harvey* collection ( ^{*hur.12*}_{0.27}) was most similar to the Archive-It *Hurricane Harvey* collection confirming our expectation that collections generated from Google may be similar to social media collections since users may use Google to discover URIs. The Twitter Moments *2016 Pulse nightclub shooting* collection ( ^{*pul.15*}_{0.34}) was third most similar even though it has no topics in common with the Archive-It *Hurricane Harvey* collection (n-gram topic similarity of 0), indicating a strong similarity across other dimensions. This shows the need for taking topic similarity into consideration before collection comparison. Similarly, the Twitter Moments collections were most similar to the Archive-It *2016 Pulse nightclub shooting* collections.

Random collections were most similar to other random collections due a common set of properties random collections show: all the random collections produced high diversity values for *Document-Term matrix* (0.93 – 0.95) and *List of Entity sets* (0.88 – 1.0) representations. Also, they included no social media sources (social media rate - 0.0) and sampled from a diverse set of hosts (hostname diversity between 0.92 – 0.77). Across the various topics, the distribution of topics (n-gram similarity) CCS metric provided the most information to distinguish the collections, producing the highest variance or spread ($\sigma = 0.29$) across the collections (Figure 23). The radar plots (Figure 35) illustrates this variance.

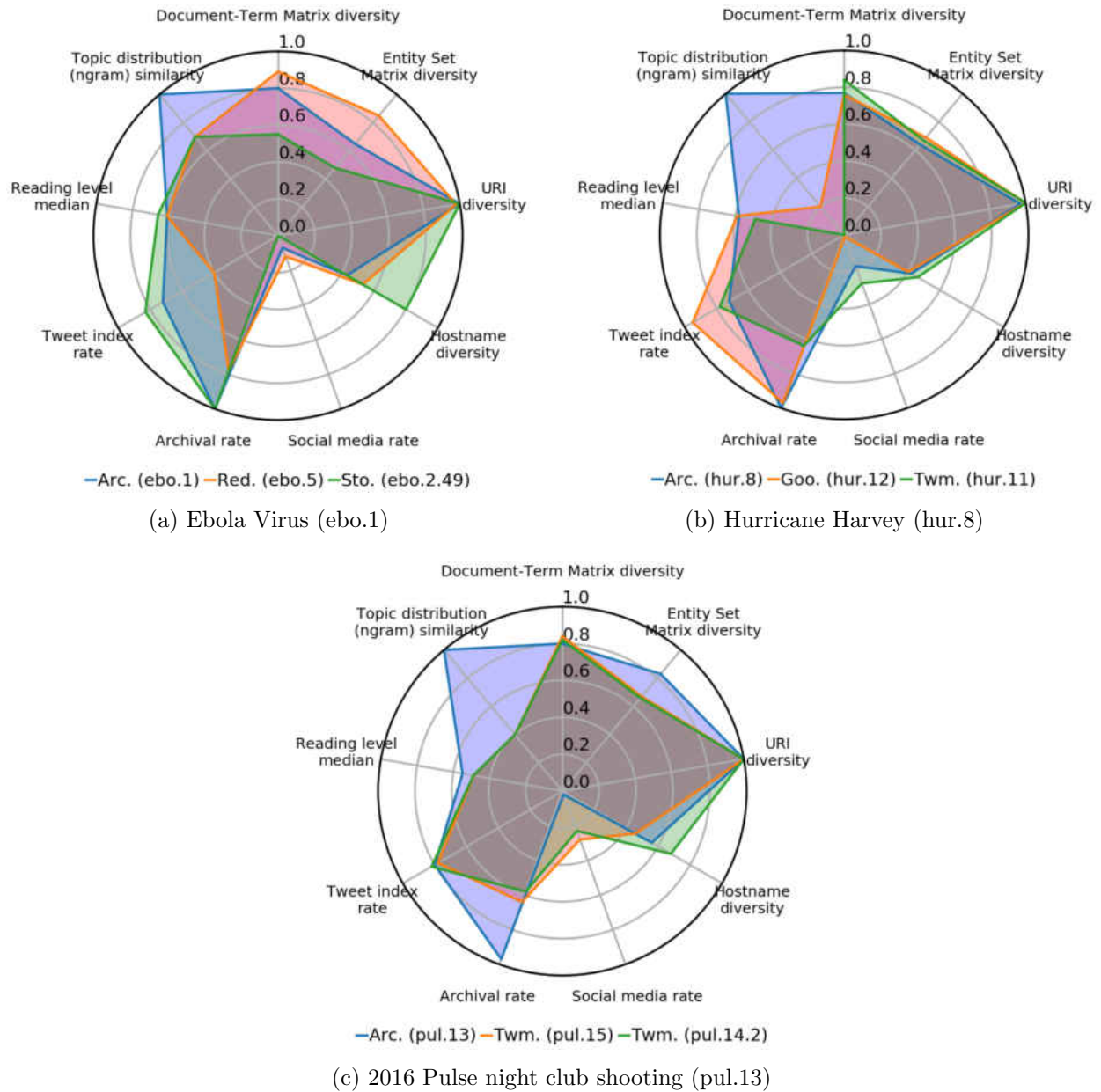


Fig. 35: Distribution of CCS Metrics for pair of collections most similar to Archive-It collections (a – c).

TABLE 23: Ranking of CCS Metrics based on the Standard Deviation of CCS values in the dataset

Rank	CCS Metric (with variants)	Standard Deviation
1	Distribution of topics	0.29
2	Hostname diversity	0.26
3	Tweet index rate	0.20
4	Archive rate	0.17
5	List of entity sets diversity	0.16
6	Social media rate	0.15
7	Document-Term Matrix diversity	0.09
8	Reading level median	0.05
9	URI diversity	0.03

This suggests the importance of collection summaries in distinguishing collections. This was followed by the hostname diversity CCS metric ($\sigma = 0.26$), suggesting multiple ways collections sample hosts. The target audience (readability) and URI diversity provided the least information to distinguish the collections: this may be explained by the idea that the documents in the collection target a common audience and have little or no duplicate links ($d_{URI} = 1.0$).

6.5 CHAPTER SUMMARY

This work introduced the Collection Characterizing Suite (CCS) as a collection of metrics that describes seeds across seven dimensions. Using the CCS, we characterized seeds generated from social media sources and compared those seeds with hand-selected seeds of experts on Archive-It. We showed that since automatically generated social media seeds were similar to hand-selected expert-generated seeds on Archive-It. Consequently, seeds from social media could be used in addition to expert-generated seeds.

CHAPTER 7

QUANTIFYING THE QUALITY OF SEEDS: QUALITY

PROXIES (QPS) FOR SEEDS

In Chapter 6 we made initial contributions to address the third research question by presenting the Collection Characterizing Suite (CCS).

- **RESEARCH QUESTION 3:** How do we evaluate automatically created collections with those generated by human experts in Archive-It?

The CCS provides a means of profiling a seed collection across seven dimensions (Chapter 6.1). Each dimension expresses a single character trait of the collection. For example, the *source diversity* (Chapter 6.1.5) expresses the level of domain variety in the collection. A pair of collections may be compared by calculating the distance between their respective CCS profile vectors.

While Chapter 6 focused on comparison, this chapter addresses the third research question by extending the comparison idea to evaluation. Specifically, we focus on how to determine the quality of the seeds in a Micro-collection. Comparison only tells us whether two collections are similar, but it does not signal if the collections are high or poor quality collections. It is insufficient to generate seeds without establishing their quality, but this raises a new challenge: “How do you define and quantify quality?” In this chapter we address this and the third research question (Chapter 7.5) by extending the idea of collection comparison to evaluation with the Quality Proxies (QPs) for seeds (Chapter 7.1) which are measures that approximate the quality of a seed.

7.1 QUALITY PROXIES (QP) FOR SEEDS

The problem of determining the quality of URIs is not new. This is the same problem search engines face when they must return a fixed list of URIs (from a candidate set of possibly millions items) to fulfill an informational request encoded in a search query. Given a search query, e.g., “ebola virus,” and 6.18 billion¹ documents, a search engine, e.g., Google,

¹Estimate of the total number of webpages according to worldwidewebsize.com as at 2020-02-19

must first determine the subset of the 47.8 million² webpages that are relevant to the query. This is an important step taken to ensure the search results are relevant. Relevance is a quality check but it is not sufficient; determining that 47.8 million documents are relevant to the query is insufficient since the user often only needs a handful of pages. Therefore, the search engine must go beyond relevance and must rank the 47.8 million candidates and select hundreds to populate the SERPs. The task of ranking done by the search engine is analogous to quantifying quality since it requires assigning a score to a URI computed from multiple metrics such as the PageRank of the URI, geographic information of user, user’s preferences, etc. In fact, according to Google [172], their search algorithm examines 200 signals (with 50 variations leading to a total of 10,000 signals) for each query. Each metric provides a means of approximating the quality of the URI. The combination of these metrics helps the search algorithm determine that webpages from sources such as `wikipedia.org`, `who.int`, and `cdc.gov` are high quality pages with respect to the “ebola virus” query and likely belong on the first page of the SERP. It is not surprising that these pages made it to the first page beating out millions of relevant webpages; they are highly popular (well-known) pages from globally-known institutions. It is not strange that relevant popular pages are more likely to be ranked higher than relevant obscure pages. In fact, this preference for popularity over obscurity is encoded in PageRank (a link analysis algorithm) which assigns higher scores to pages with a large number of inlinks, in other words, to popular pages. In summary, search engines use *popularity* as one method to approximate quality. **Popularity is a proxy for quality**, or can be considered a Quality Proxy (QP). This is reasonable since one can argue that popularity is the reward for quality. Therefore, we may approximate the quality of a webpage by measuring how popular it is. There may be exceptions for which popularity does not mean quality [173, 174], but that is beyond the scope of our work. It is also important to note that popularity is a function of time. For example, following the 2019 Democratic Presidential campaign season, Democratic candidate Pete Buttigieg added 1.3 million followers with a growth rate of 1,350% [175]. Twitter is the primary social media platform we utilized to illustrate the QPs, however, the QPs are applicable to other social media platforms, such as Facebook and Reddit.

In this work, we utilize popularity as a quality approximation for seeds. In the next section, we explore it as a Quality Proxy for seeds and also argue that popularity is not sufficient. In Chapter 7.3 we explore additional non-popularity based QPs. The Quality Proxies that determine the quality of seed URIs extracted from Micro-collections can be

²Number of hits the query “ebola virus” returned from the Google SERP on February 14, 2020.

TABLE 24: Summary of the Quality Proxies (QPs) for seeds.

#	Quality Proxy	Group	Section
1	Post	Popularity	7.2
2	Author	Popularity	7.2
3	Domain	Popularity	7.2
4	Geographical	Proximity	7.3.1
5	Temporal	Proximity	7.3.2
6	Subject expert	Proximity	7.3.3
7	Retrievability	Proximity	7.3.4
8	Relevance	Proximity	7.3.6
9	Reputation	Uncategorized	7.3.5
10	Scarcity	Uncategorized	7.3.7

grouped into *popularity*, *proximity*, and *uncategorized*, as summarized in Table 24.

7.2 POPULARITY SEED QUALITY PROXY

There are generally two approaches toward quantifying the popularity of URIs. The first approach, the link-based approach [47] utilizes the link structure of the Web to assign weights to webpages. PageRank is a link-based approach. The second approach leverages social media post statistics to assign popularity scores to URIs embedded in social media posts. Social media posts often keep statistics that track the number of times a post is shared (a “retweet” on Twitter), liked, or replied to. Transitively, the popularity of URIs embedded in a social media posts can be derived from the social media post statistics [176, 177, 178]. The popularity of a seed can also be derived from the popularity of the user who posted the seed. Since the link-based approach of determining popularity is computationally expensive because it requires crawling the Web, we utilize social media post statistics to assign popularity scores to URIs embedded in the social media posts, in addition to the popularity of the user that posted the seed. Our popularity Quality Proxy uses a five-dimensional vector to capture the popularity of seed (Figure 36). The first three dimensions (*likes*, *shares*, and *replies*) express the popularity of the social media post. The fourth (*author popularity*) expresses the popularity of the author of the social media post. The last (*domain social media popularity*) expresses the popularity of the social media account

associated with the seed domain.

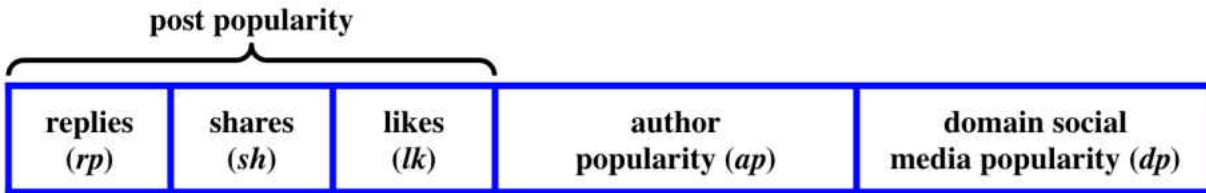


Fig. 36: Five-dimensional vector expressing the popularity of a seed embedded in a social media post.

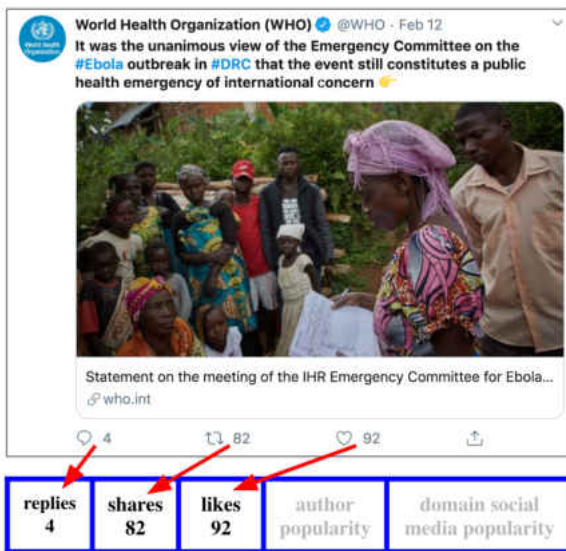
7.2.1 THE SOCIAL MEDIA POST POPULARITY DIMENSIONS

The first three dimensions of the Quality Proxy vector express the popularity of the URI embedded in a post: how many people replied (*replies rp*), shared (*shares sh*), and liked (*likes lk*) the social media post. All of these post statistics are normalized ($x_{normalized} = \frac{x - \min_X}{\max_X - \min_X}$) before populating the vector. Figure 37 illustrates the population of the *post popularity* dimensions of two tweet popularity vectors. The *post popularity* dimensions can be additionally populated with k social media posts that embed the seed.

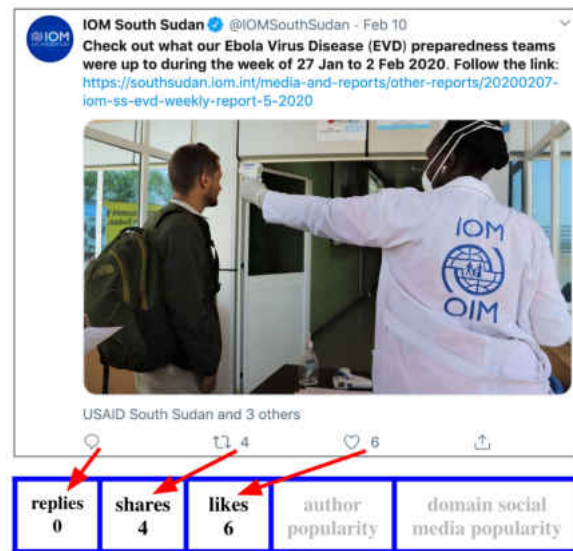
7.2.2 THE AUTHOR POPULARITY DIMENSION

The fourth dimension, *author popularity ap*, expresses the popularity of the author who created the social media post. Social media sites often have statistics to quantify how popular a social media account is. For example, Twitter and Instagram count *followers* (in-degree or incoming links), the number of people following an account, and *following* or *friends* (out-degree or outgoing links), the number of accounts a user follows. Unlike Twitter, which separately counts in-degree (*followers*) and out-degree (*following*), Facebook only counts *friends* (in-degree and out-degree). A Facebook *friend* expresses a bi-directional relationship.

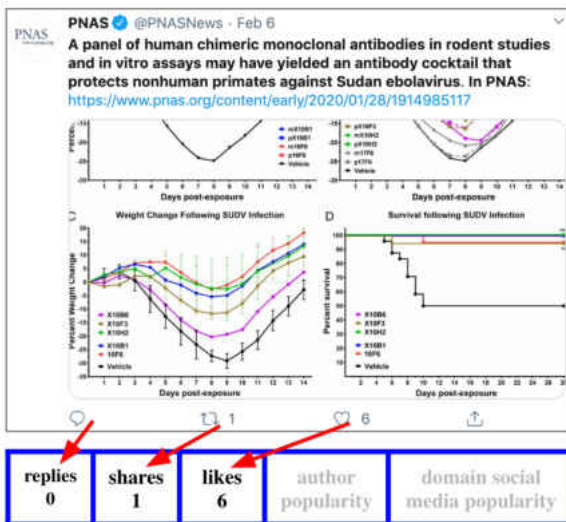
For social media platforms like Facebook with bi-directional links, the author popularity metric ap is simply the normalized count of this metric (e.g., *friends*). For social media platforms like Twitter, $ap = \text{in-degree} - \text{out-degree}$ (e.g., *followers - following* for Twitter) normalized. If the in-degree $<$ out-degree, then $ap < 0$. To fix this, the offset (the absolute value of smallest difference between in-degree and out-degree) is added to each difference before normalization. Given a set of social media posts P , let in_i and out_i represent the in-degree and out-degree of social media post i , respectively. There are multiple studies



(a) Popularity vector of seed 1



(b) Popularity vector of seed 2



(c) Popularity vector of seed 3



(d) Popularity vector of seed 4

Fig. 37: Population of the *post popularity* dimensions of four seed popularity vectors from the *replies*, *likes*, and *shares* statistics of their respective containing tweets. The *post popularity* dimensions can be additionally populated with k social media posts that embed the seed.

[176, 177, 178] that have utilized the content or tweet statistics to rank tweets. The *author popularity*, ap is given by Equation 10 which is similar to *FollowerRank* from Nagmoti et al. [178] ($\frac{in}{in+out}$), but sensitive to the magnitude of followers when considering two users with the same ratio of followers/following. For example, unlike ap , *FollowerRank* assigns the same score (0.67) to a given user with 20 followers/10 following, and another user with 20,000 followers/10,000 following. Table 25 outlines the calculation of ap for four seeds extracted on February 15, 2020. The *offset* is 0.0 since the minimum d_i , $3,723 \geq 0$. The difference between the minimum and maximum d_i , $(5,394,414 = 5,398,137 - 3,723)$ was used to normalize ap_i . The dp_i value is calculated in the same fashion as ap_i with one important difference: the *in* and *out*-degree information is extracted from the Twitter handle that has a bi-directional (Figure 38) link with the domain of the seed.

$$ap_i = \frac{d_i + offset}{\max_{i \in P}(d_i) - \min_{i \in P}(d_i)}$$

$$d_i = in_i - out_i \tag{10}$$

$$offset = \begin{cases} 0 & ; \text{if } \min_{i \in P}(d_i) \geq 0 \\ |\min_{i \in P}(d_i)| & ; \text{otherwise} \end{cases}$$

Populating the *author popularity* dimensions of seeds requires knowing the in-degree and out-degrees of the respective post authors. Let us consider populating the author popularity vectors of the four seeds in Figure 37:

- **Seed 1:** [https://www.who.int/news-room/detail/12-02-2020-statement-on-the-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-for-ebola-virus-disease-in-the-democratic-republic-of-the-congo-on-12-february-2020](https://www.who.int/news-room/detail/12-02-2020-statement-on-the-meeting-of-the-international-health-regulations-(2005)-emergency-committee-for-ebola-virus-disease-in-the-democratic-republic-of-the-congo-on-12-february-2020)
- **Seed 2:** <https://southsudan.iom.int/media-and-reports/other-reports/20200207-iom-ss-evd-weekly-report-5-2020>
- **Seed 3:** <https://www.pnas.org/content/early/2020/01/28/1914985117>
- **Seed 4:** <https://www.labroots.com/trending/microbiology/16778/ebola-outbreak-continues-researchers-create-faster-genetic-test>

Table 25 illustrates the computation of the *author popularity* ap for the four seeds from Tweets.

TABLE 25: The ap_i values of four seeds from Figure 37. The in and out -degree details were extracted on February 15, 2020. The $offset = 0$ since the minimum d_i , $3,723 \geq 0$. The difference between the minimum and maximum d_i , $(5,394,414 = 5,398,137 - 3,723)$ was used to normalize ap_i . dp_i is calculated in the same fashion as ap_i with one important difference: the in and out -degree information is extracted from the Twitter handle that has a bi-directional (Figure 38) link with the domain of the seed.

$Seed_i$	Author	in-degree (in_i)	out-degree (out_i)	$d_i = in_i - out_i$	ap_i (Eqn. 10)
1	@WHO	5,399,854	1,717	5,398,137	1
2	@IOMSouthSudan	9,237	740	8,497	0.0008
3	@PNASNews	118,866	1,343	117,523	0.0210
4	@Microbiology_LR	3,886	163	3,723	0
			$\min_{i \in Posts} (d_i)$	3,723	
			$\max_{i \in Posts} (d_i)$	5,398,137	

7.2.3 THE DOMAIN POPULARITY DIMENSION

The *domain social media popularity* dp metric quantifies the popularity of the seed domain, instead of utilizing statistics (*likes*, *shares*, and *replies*) found in the post to assign popularity. This metric attempts to approximate the popularity of the seed domain as opposed to the popularity of the author who posted the seed (ap). The popularity of a seed domain is extracted from the in and out -degree information of the social media account of the domain. It is akin to the PageRank score of the domain, but since we do not crawl the web, we utilize social media to approximate the popularity of a seed domain. For example, still utilizing Twitter as our example, @WHO posted the first seed (Figure 37) from the domain, `who.int`. To calculate dp for the first seed, first, we must find the social media account (<https://twitter.com/WHO/>) associated with the `who.int` domain. Second, we extract the in - and out -degree details from the account. Third, we apply the same method for calculating ap (Chapter 7.2.2) to calculate dp .

The challenge in calculating dp is finding the social media account associated with a domain which we address as follows. First, we extract all links from the front page of the domain (e.g., <https://www.who.int/index.html> for `who.int`). Second, we identify all

Twitter handles from the front page. If no handles are found from the front page, we issue a Google search with query “domain twitter handle” (e.g., “who.int twitter handle”) and extract all Twitter handles returned from the first page of the Google SERP. For domains with multiple Twitter handles, all handles from the second step are collected. Third, for all Twitter handles selected in the previous step, we check if there exists a single handle that points to the respective domain to verify a bi-directional link. We want to ensure that the domain links to the Twitter account and that the Twitter account links to the domain. For example, the Twitter profile of @WHO includes a link with the who.int domain, and the who.int domain likewise links to the @WHO Twitter profile. This establishes a bi-directional linkage (Figure 38). For a given domain, if a bi-directional linkage cannot be verified, the dp for the domain is set to zero.

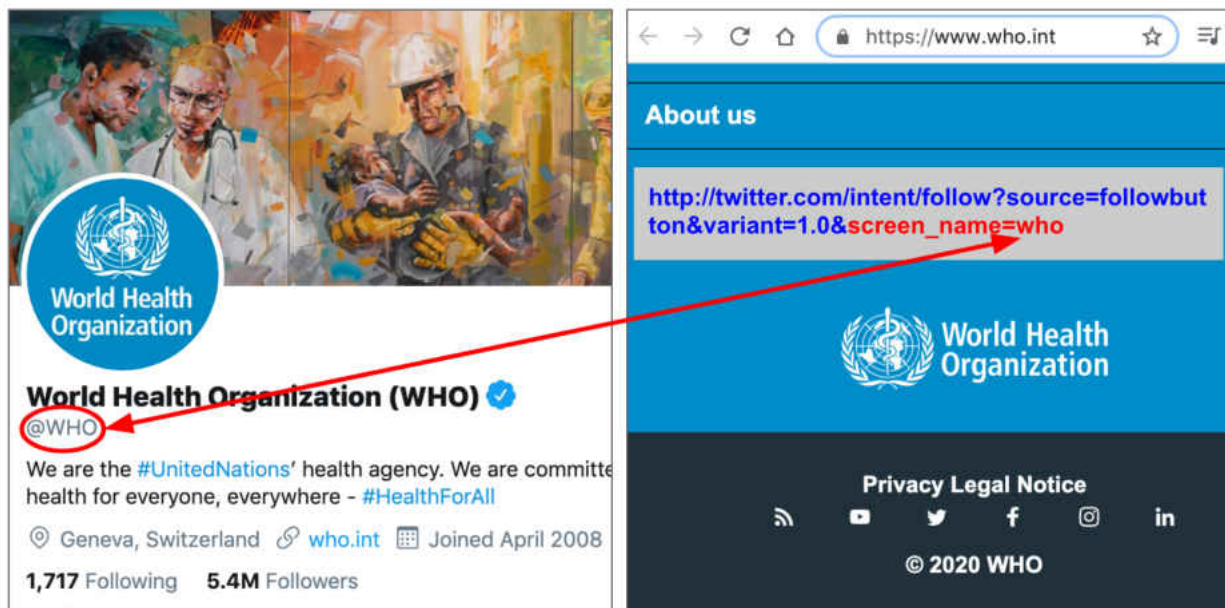


Fig. 38: An illustration of a bi-directional link; the Twitter account @WHO (left) links to the who.int front page (right), and the who.int front page (right) links to the @WHO (left) Twitter account. The presence of a bi-directional link validates that the who.int domain is associated with @WHO Twitter handle, and the use of @WHO as a source for the *in* and *out*-degree information needed to calculate dp . The screenshot on the right has been edited to show more detail.

7.3 NON-POPULARITY SEED QUALITY PROXIES

Popularity Quality Proxies favor social media posts from popular accounts or posts. Let

us consider two seeds about the *Flint Water Crisis* story to illustrate the effects of the popularity Quality Proxies on two seeds, the first from a large international news organization (CNN) and the second from a smaller local news organization (MLive). A seed tweeted by @CNN (about 45 million followers as of 2020-02-19) would score higher than a seed from @MLive (about 300,000 followers as of 2020-02-19) across the $ap = dp$ dimensions. A seed from CNN also has a higher likelihood of being replied to, shared, and liked than a seed from MLive because of the larger audience size of CNN. While the popularity method of assigning quality is less likely to give credit to URIs from spam accounts, it is flawed for multiple reasons. First, account popularity can be artificially manipulated. For example, one could purchase followers [179, 180] in order to boost the authority of an account. Second, not all authoritative sources are popular. For example, the MLive local media organization is located in Michigan along with the city of Flint, the epicenter of the *Flint Water Crisis*. Consequently, one could argue that MLive is a local authority on topics about the *Flint Water Crisis*, more so than CNN, a national and international news organization. In fact, according to Denise Robbins, it took the national media one year after the E. coli outbreak to report the Flint story [31]. Consequently, it is pertinent to quantify authority across other dimensions in addition to the popularity dimension. This is the rationale for the following non-popularity based Quality Proxy metrics (Table 24, No. 4 – 10, Figure 39): *Geographical*, *Temporal*, *Subject expert*, *Retrievability*, *Relevance*, *Reputation*, and *Scarcity*. The proximity Quality Proxy metrics assign a seed a quality score based on some notion of proximity.

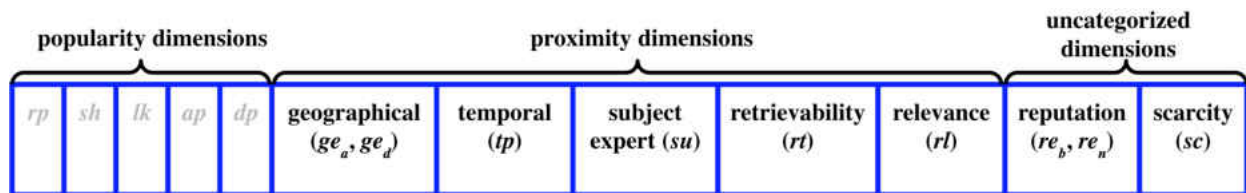


Fig. 39: Seven additional non-popularity (proximity and uncategorized) dimensions of the seed authority vector expressing quality of the seed across *geographical* (author and domain), *temporal*, *subject expert*, *retrievability*, *relevance*, *reputation* (broad and narrow), and *scarcity* dimensions.

7.3.1 GEOGRAPHICAL QUALITY PROXY

The stories and events for which we generate seeds are often associated with some geographical location. For example, the *2014 Ebola virus outbreak* primarily affected the Western African countries of Guinea, Liberia, and Sierra Leone. The epicenter of the *Flint Water Crisis* was Flint, Michigan. *Hurricane Harvey* made landfall in Texas and Louisiana in August 2017. The rationale for the *geographical ge* Quality Proxy metric is to assign credit to a local source (local authority) when we have prior knowledge about the geographical location associated with a story or event. The local source could be an individual (ge_a - author geographical QP) or an organization (ge_d - domain geographical QP). For example, given two seeds from CNN (national/international media) and MLive (Michigan local media), the ge_d QP metric would give more credit to the MLive (`mlive.com`) seed since MLive is closer to Flint, Michigan. Similarly, given two individuals, a resident of Flint, Michigan, and a resident of San Francisco, California, the ge_a would give more credit to the Flint resident.

We assign the ge_a QP by first extracting the geographical information of the post author if available. For example, from Figure 38 (left image), the Twitter account of @WHO is tagged with the location Geneva, Switzerland (46.2044° N, 6.1432° E). Given the event, *2014 Ebola virus outbreak* with epicenter in West Africa (13.5317° N, 2.4604° W), posts (e.g., tweets) from @WHO are assigned an initial geographical proximity score of the distance in miles (about 2,267 miles) of the two coordinates as measured by the Haversine formula (Equation 11). We normalized both the ge_a and ge_d distances ($[0, 1]$) to permit performing vector operations with other non-geographical components of the seed QP vector.

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

r : radius of sphere (Earth)

d : distance between the two points along a great circle of the sphere (KM) (11)

φ_1, φ_2 : latitude of point 1 and latitude of point 2 (in radians)

λ_1, λ_2 : longitude of point 1 and longitude of point 2 (in radians)

The ge_d QP is assigned from the social media account associated with the seed domain. The same method employed to find the account to derive dp is used for ge_d .

The ge_a QP score relies on deriving geographical locations associated with the author of a social media post. This is challenging for multiple reasons. First, not all authors provide this information. Such authors are assigned $ge_a = 0$. Second, the geographical information might be expressed in a non-machine easily-readable form, e.g., “UK & New York.” While a human can easily understand that this location combines two places, a machine without the application of some special location segmentation algorithm would erroneously parse

the string as a single location. Third, the same locations can have different names. For example, New York City can be expressed in different ways: “New York City,” “New York,” “NYC,” and “New York, New York.” To address this problem we utilized the Google Maps Services Places API [181] to normalize the names of locations into a single name and its corresponding geo-coordinates.

7.3.2 TEMPORAL QUALITY PROXY

The stories and events for which we generate seeds often happen at a place (or places), but always happen at some time. After the occurrence of the event or before its occurrence, news organizations report the story or event. For example, some of the earliest reports of the *Flint Water Crisis* story are from `mlive.com`. The temporal Quality Proxy tp was chosen to assign a Quality Proxy to a seed published “early.” When events (especially long-running) become popular, it is more likely for many different news organizations to cover the event. However, a seed created during the initial stages of the story, when it was not popular, deserves credit especially since it could contain contemporary developments.

The challenge in using the temporal Quality Proxy is the determination of what is “early.” We consider this information subjective and thus only use this Quality Proxy when apriori information about what constitutes early is present. In such cases, similar to the geographical Quality Proxy metric, the time difference is calculated between the publication date of the seed and the reference point considered early. The difference is subsequently normalized before placement into the seventh-dimension of the seed Quality Proxy vector.

7.3.3 SUBJECT EXPERT QUALITY PROXY

The subject expert Quality Proxy su attempts to reward subject expertise to the domain of a seed. For example, given two seeds about the *Ebola virus outbreak*, one from the Centers for Disease Control and Prevention (CDC), and another from the blog of a high school senior, we would expect to assign the CDC a higher subject expert quality score since the CDC is an authority on health topics.

While the idea for the assigning su scores is easy to explain and justify, it is a difficult task to measure the subject expertise of a seed. How does one measure the subject expertise of `cdc.gov`? To address this problem we posit the following: *A subject expert often has more to say about their subject of expertise.* This means, if indeed the CDC is an expert on Ebola, we would expect to see many more reports from the CDC about Ebola than from others.

TABLE 26: Illustration of the assignment of *su* scores for two seed domains (*cdc.gov* vs. *espn.com*) for the query “ebola virus.”. We use the count of result pages from the Google SERP (Figure 40) to estimate the *subject expertise* of the domain of a seed. Accordingly, *cdc.gov* has a higher subject expertise than *espn.com*. These counts were derived from queries issued on February 16, 2020.

Statistic	<i>cdc.gov</i>	<i>espn.com</i>
Number of pages in website	951,000	11,100,000
Number of pages in website with query “ebola virus”	15,800	152
Seed domain <i>subject expert su</i> score	$\frac{15,800}{951,000} = 0.016614$	$\frac{152}{11,100,000} = 0.000013$

We acknowledge that this is a simplifying assumption that could be exploited. We used the Document Frequency (DF) to approximate the subject expertise of the domain of a seed. We extract DF scores by counting the number of result pages (e.g., Figure 40) returned by the Google SERP for a given query normalized by the total number of pages indexed by the search engine for the site. The normalization is required in order to avoid giving more advantage to larger websites. The DF simply counts the number of documents from a domain (e.g., *cdc.gov*) that has a particular term (e.g., *ebola virus*).

Table 26 illustrates the assignment of *su* scores with respect to *ebola virus* for two domains, *cdc.gov* and *espn.com*. The *su* score for a seed’s domain is simply calculated by dividing the number of pages in a website with the query by the number of pages in the website. Both values are extracted by counting the number of result pages (Figure 40). For *cdc.gov* the *Number of pages in website* extracted by issuing the “site:cdc.gov” query to Google was 951,000 as of February 16, 2020. While the *Number of pages in website with the query term “ebola virus,”* extracted by issuing the “ebola virus site:cdc.gov” query to Google was 15,800. *espn.com* has a larger website (11,100,000 webpages), but only a small fraction of the website pages have the “ebola virus” term (152 webpages). Therefore, according to the *su* Quality Proxy metric, *cdc.gov* ($su_{cdc.gov} = 0.016614$) is more of a subject expert than *espn.com* ($su_{espn.com} = 0.000013$) for the *Ebola virus* subject.

7.3.4 RETRIEVABILITY QUALITY PROXY

The *retrievability rt* QP metric estimates how easy a seed is to find. Extracting seeds

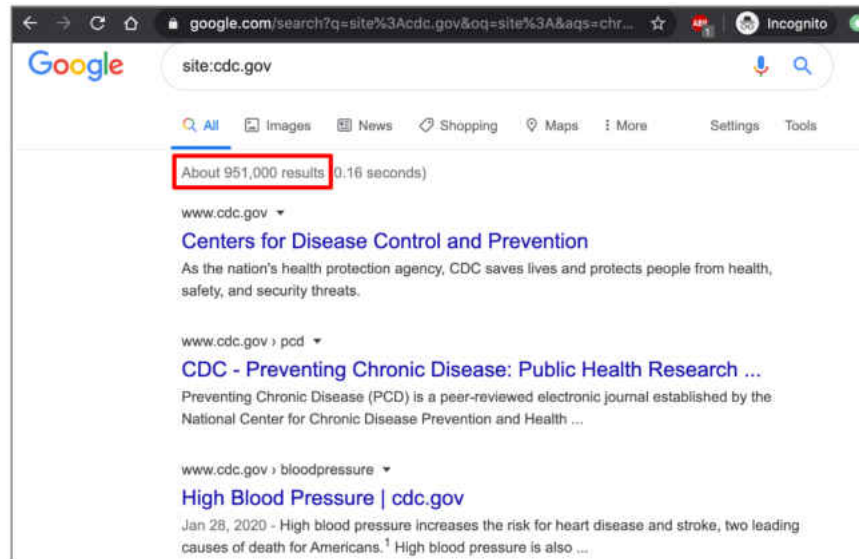


Fig. 40: Google SERP showing (red annotation) the count (951,000) of result pages for query: “site:cdc.gov” - an estimate of the total number of pages from `cdc.gov` indexed by Google. We use this statistic from the SERP to calculate (Table 26) the *su* score for a seed.

from Micro-collections requires more effort than scraping Web search engine SERPs. For example, generating a collection of URIs of the $\mathbf{P}_n\mathbf{A}_1$ or $\mathbf{P}_n\mathbf{A}_n$ post class requires independently dereferencing each social media post and extracting the replies from the post. Therefore, if the URIs discovered from Micro-collections are easily discoverable via a search engine such as Google, it does not justify the extra effort of extracting seeds from Micro-collections. For this reason, the *rt* Quality Proxy metric quantifies the level of difficulty of finding a seed. For example, Wikipedia pages for various entities (e.g., political figures) are often placed on the front page of SERPs, meaning they have high retrievability. It is therefore a desirable quality to identify relevant seeds that are not easy to find.

Azzopardi and Vinay’s [182] retrievability measure (Equation 12) quantifies how a retrieval system affects the users’ ability to access information. The retrievability $r(d)$ of a seed d given by Equation 12 counts the number of queries $q \in Q$ that successfully find a document d at a rank lower threshold c . Unlike Azzopardi and Vinay’s concern of measuring the retrievability of a system, we measured the retrievability of individual documents similar to Traub et al. [183]. We approximated *rt* of a seed with the reciprocal rank $\frac{1}{rank_d}$ when searching the first k Google SERPs for the seed with the query used to extract seeds (e.g., “ebola virus”). This means that the *rt* of a seed is a function of the time that *rt* was

measured.

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, c) \quad (12)$$

7.3.5 REPUTATION QUALITY PROXY

The seeds we extract from Micro-collections in social media originate from various sources with varying reputations. Given two seeds about “ebola virus,” one from InfoWars (known to promote conspiracy theories [184]) and another from the CDC, it is clear that it would be problematic to consider the quality of information derived from both sources as equal instead of attributing CDC as the higher quality source. Similar to the *subject expert* Quality Proxy, the *reputation re* Quality Proxy metric attempts to attribute reputation to seeds (assigned to their domain).

We defined two kinds of reputation, broad - re_b and narrow - re_n . Broad reputation attributes reputation to the domain of a seed for having a record of publishing content about a topic, while narrow reputation attributes reputation to the domain of a seed for having a record of publishing content focused specifically on a story. For example, `cdv.gov` does not only report about Ebola virus, it has a reputation for publishing on multiple *health* topics. In contrast, the only focus of `ebolafacts.com` is publishing content focused primarily on Ebola virus. Therefore, `cdc.gov` has both a broad reputation (for *health* topics) and a narrow reputation (for *Ebola virus*). However, `ebolafacts.com` does not have a broad reputation for health topics, but has a narrow reputation for *Ebola virus*. But the question remains, how does one approximate reputation³? We addressed assigning *re* scores to seeds by leveraging the expertise of Wikipedia editors. We posit that *Wikipedia editors presumably sample reputable sources*. Specifically, the reputation of the domain of a seed corresponds to the fraction of times it was cited as a reference from a gold-standard set of Wikipedia articles.

For re_b , the gold-standard is represented by a collection of Wikipedia articles that focus on the topic (e.g., *Disease outbreaks*) of the seed. For re_n , the gold-standard is represented by the canonical Wikipedia page for the story. The canonical page can be found by searching for the top ranked Wikipedia page for the query (e.g., “ebola virus outbreak”) representing the topic. To assign re_b or re_n to the domain of a seed, we extracted the URIs from the references of the reputation gold-standard Wikipedia articles and calculated the fraction of times each domain was referenced. For example, in our reputation gold-standard for the

³It is important to note that attributing reputation is not the same as attributing political orientation.

*Disease outbreaks*⁴ topic, `cdc.gov` appeared 42 out of 57 gold-standard articles. Therefore, the `cdc.gov` domain has a re_b score of 0.74. The `cdc.gov` domain appears 14 times out of 720 references in the canonical *2014 Western African Ebola Virus Outbreak* Wikipedia⁵ page, and thus has a 0.02 re_n score.

7.3.6 RELEVANCE QUALITY PROXY

The *relevance* rl QP measures the degree to which a seed is on-topic. A seed that receives high marks across all the other QP dimensions remains non-relevant if it is off-topic. We approximate relevance by simply measuring the cosine similarity between a seed’s document vector and a gold-standard document vector that captures our definition of relevance. The gold-standard is created by concatenating the text of hand-selected documents that are relevant to a topic, and creating a feature (vocabulary) vector consisting of the TF or TFIDF weights of the terms in the concatenated document.

7.3.7 SCARCITY QUALITY PROXY

The *scarcity* sc QP rewards seeds from domains that are rare in a collection of seeds. It is not surprising to find multiple seeds from news organizations (e.g., `cnn.com`, `foxnews.com`, `bbc.co.uk`) for news topics. Sometimes far-reaching news events are covered by organizations for which news is not their primary domain (e.g, `eonline.com` and `espn.com`) and which may offer a novel reporting perspective. The sc QP was created to surface such seeds and is approximated by $1 - \frac{|d_s|}{N}$, where d_s is the frequency of a seed’s domain out of N total domains.

7.4 ADDITIONAL QUALITY PROXIES: FLIPPING QUALITY PROXIES

Thus far, the Quality Proxies have been presented with the assumption that the higher the QP value, the better the trait the QP captures. For example, a high author popularity ap score is a desirable trait, and a low author popularity score is not a desirable trait. However, desirability can be subjective. This means a curator might desire to surface seeds from authors that are not popular in an effort to amplify the voices of obscure users. Consequently, this requires flipping the direction of the reward system of QP under consideration. For example, before flipping, the most popular author would have $ap = 1$, but if we flipped

⁴https://en.wikipedia.org/wiki/List_of_epidemics

⁵https://en.wikipedia.org/wiki/Western_African_Ebola_virus_epidemic

(represented with bar over the QP) the ap Quality Proxy, $\overline{ap} = 0$ is assigned to the most popular author. Since all the quality proxies were designed to fall within $[0, 1]$, a QP qp is simply flipped by $1 - qp$; $\overline{qp} = 1 - qp$.

The ability to flip QPs provides us with additional QPs (\overline{rp} , $\overline{ge_a}$, \overline{rt} , etc). But it must be noted that the unflipped (qp) state and the flipped (\overline{qp}) state of QPs are mutually exclusive: a switch cannot be ON and OFF at the same time.

7.5 THE SEED QUALITY PROXY MATRIX AND COMPARING SEEDS

$seed_1$	rp_1	sh_1	lk_1	ap_1	dp_1	ge_{a1}	ge_{d1}	tp_1	su_1	rt_1	rl_1	re_{b1}	re_{n1}	sc_1
$seed_2$	rp_2	sh_2	lk_2	ap_2	dp_2	ge_{a2}	ge_{d2}	tp_2	su_2	rt_2	rl_2	re_{b2}	re_{n2}	sc_2
$seed_3$	rp_3	sh_3	lk_3	ap_3	dp_3	ge_{a3}	ge_{d3}	tp_3	su_3	rt_3	rl_3	re_{b3}	re_{n3}	sc_3

Fig. 41: A seed Quality Proxy matrix \mathbf{Q} . Each row represents a 14-dimensional seed Quality Proxy vector \mathbf{q} for a seed $seed_i$

The seed Quality Proxy vector \mathbf{q} is a 14-dimensional vector ($\mathbf{q} \in \mathbb{R}^{14}$) of all the Quality Proxies introduced thus far. The dimension may be increase when new QPs are found. The metrics individually express the quality of a seed across 14 different dimensions. Each metric $q_i \in \mathbf{q}$ is normalized ($q_i \in [0, 1]$) and designed such that 0 represents worst quality and 1 represents best quality. Multiple QP vectors make up the seed quality matrix $\mathbf{Q} \in \mathbb{R}^{n \times 14}$.

A user can control the relative importance of the metrics of \mathbf{q} depending on prior information or specific needs. Therefore, one can multiply a weight vector $\mathbf{w} \in \mathbb{R}^{14}$ ($\sum_{i=1}^n w_i = 1$) with \mathbf{q} to reflect the importance of each metric to obtain a new Quality Proxy scores q'_i ($q'_i = q_i w_i$). The weight vector can also be used to switch off specific metrics. For example to switch off q_i , we set $w_i = 0$, such that $q_i w_i = 0$.

Since the quality of a seed is encoded as a vector \mathbf{q} and the quality of a collection of seeds is encoded as a matrix \mathbf{Q} , seeds can be compared through vector operations and collections can be compared through matrix operations.

The Quality Proxy score q of a seed can be instantiated by the norm of the n-dimensional Quality Proxy vector \mathbf{q} of the seed.

$$q = |\mathbf{q}|_2 = \sqrt{\sum_{i=1}^n q_i^2} \quad (13)$$

Similarly, the Quality Proxy score Q of a collection can be instantiated by the $m \times n$ -dimensional norm of the Quality Proxy matrix \mathbf{Q} of the collection

$$Q = \|\mathbf{Q}\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^n q_{ij}^2} \quad (14)$$

To compare two seeds (a and b) or two collections (A and B), we compare the seed (q_a vs q_b) or collection (q_A vs q_B) QP scores. In the next chapter, we will demonstrate how Quality Proxies independently behave like alphabets that can be combined in different ways to provide various policies for generating seeds.

7.6 CHAPTER SUMMARY

In this chapter we addressed the third research question by going from comparison to evaluation. This was done by presenting the seed Quality Proxies that express the quality of a seed across popularity-based and non-popularity based dimensions. While popularity assigns quality scores based on the popularity of the post containing the seed or the popularity of the seed's domain, proximity-based metrics attributes quality based on some notion of proximity. Collectively, the Quality Proxies make up the seed Quality Proxy vector, and enable assigning quality scores to seeds and/or quality scores to collections, thereby facilitating the comparison and evaluation of seeds and collections.

CHAPTER 8

EXPLORING COLLECTIONS WITH QUALITY PROXIES

In the previous chapter, we introduced the Quality Proxies (QPs). A single Quality Proxy qp , flipped (\overline{qp}) or unflipped (qp) independently attributes some quality trait to a seed. For example, a high ge_a gives credit to a seed posted by an author close to the geographical region (epicenter) associated with the seed topic, while $\overline{ge_a}$ gives credit to a seed posted by a distant author. The ge_d Quality Proxy is similar to ge_a but focuses on the seed domain's¹ proximity to the geographical epicenter instead of the author. Chapter 7.3.1 outlines how to extract ge_a and ge_d . Combining the scores from ge_a (or $\overline{ge_a}$) and ge_d (or $\overline{ge_d}$) results in new Quality Proxies with high values that can be interpreted as follows, considering *Flint Water Crisis* as the example seed topic:

- ge_a and ge_d : The author (e.g., Michigan native, Michael Moore) and seed domain (`mlive.com`) are close to the geographical epicenter (e.g., Flint, Michigan) of the seed topic (*Flint Water Crisis*). In other words, the author is a local and the domain belongs to a local news source.
- ge_a and $\overline{ge_d}$: The author is a local, but the domain is not a local source (`bbc.com`).
- $\overline{ge_a}$ and ge_d : The author is not a local (e.g., Los Angeles resident, Alyssa Milano), but the domain is from a local source (`detroitnews.com`).
- $\overline{ge_a}$ and $\overline{ge_d}$: Both author (e.g., BBC) and domain (`bbc.com`) are not local.

These four combinations of Quality Proxies show us that Quality Proxies independently behave like alphabets that can be combined in different ways to produce words that represent different ideas.

In this chapter, we explore additional combinations of Quality Proxies and study what they indicate about seeds from Micro-collections for the following stories and events: *The 2020 Coronavirus Pandemic* (Chapter 8.1), *The Flint Water Crisis* (Chapter 8.2), and *Hurricane Harvey* (Chapter 8.3). For each, we study the nature of seeds selected when scored with different combinations (e.g., rp , sh , lk) of Quality Proxies. The scores, calculated

¹More specifically the organization associated with the seed's domain.

according to Equation 13, are used to rank the seeds. We will present the top five seeds for each topic.

8.1 THE 2020 CORONAVIRUS PANDEMIC

This section references Tables 27, 28 and 29, which report on seeds sampled from a collection of 573 seeds extracted on 2020-04-09 from Micro-collections from the Twitter-Top vertical with the “coronavirus” query.

Table 27 illustrates that a combination of popularity-based Quality Proxies rp , sh , lk unsurprisingly gives more credit to seeds from popular (well-known) domains (e.g, `reuters.com`, `cnbc.com`, `gov.uk`, `washingtonpost.com`, and `wsj.com`) posted by popular authors (e.g, `@HillaryClinton`, `@CNBC`, and `@SenSanders`). Seeds from well-known domains

TABLE 27: For *The 2020 Coronavirus Pandemic*, top five seeds extracted by combining three popularity-based QPs rp , sh , lk to produce a single QP score (q - Equation 13), ranking the seeds by their QP scores, and selecting the top five seeds with the highest scores. The table illustrates how popularity-based Quality Proxies unsurprisingly gives more credit to seeds from popular (well-known) domains.

#	domain: title (author’s twitter handle)	QP Normalized				QP (Thousands)		
		q	rp	sh	lk	rp	sh	lk
1	<code>reuters.com</code> : Most Americans, unlike Trump, want mail-in ballots for November if coronavirus threatens: Reuters/Ipsos poll (<code>@HillaryClinton</code>)	1.00	1.00	1.00	1.00	13.40	31.68	101.2
2	<code>cnbc.com</code> : Chamath Palihapitiya: US shouldn’t bail out hedge funds, billionaires (<code>@CNBC</code>)	0.54	0.26	0.67	0.59	3.484	21.32	59.91
3	<code>gov.uk</code> : New immigration system: what you need to know (<code>@nicktolhurst</code>)	0.39	0.56	0.23	0.30	7.540	7.198	30.05
4	<code>washingtonpost.com</code> : When coronavirus hits, but the water is shut off (<code>@SenSanders</code>)	0.32	0.08	0.32	0.46	1.086	9.992	46.46
5	<code>wsj.com</code> : Trumps Wasted Briefings (<code>@TheRickWilson</code>)	0.25	0.15	0.28	0.29	2.069	8.966	29.83

are more likely to be replied to (*rp*), shared (*sh*), or liked (*lk*) as a result of the large audience they enjoy. The average (as of 2020-05-04) number of followers of the top five users in Table 27 is 12.08 million, causing the top five seeds to be replied to by an average of 5,515 users, shared by an average of 15,831 users, and liked by an average of 53,490 users.

The attention a seed receives is further amplified when it is posted by a well-known author. For example, 2016 US Democratic presidential candidate Hillary Clinton (@HillaryClinton with 27.6 million followers on 2020-05-04) posted the first seed, so it comes as no surprise that her popularity boosted its visibility. Similarly, 2016/2020 Democratic presidential candidate Senator Bernie Sanders (@SenSanders with 9.6 million followers on 2020-05-04), posted the fourth top seed. Sampling seeds from popular sources could help reduce spam or reduce the number of non-credible sources.

Unlike Table 27, Table 28 shifts the reward system of seeds by prioritizing authors (ge_a) and domains (ge_d) geographical close (Section I and III) or distant (Section II) to different epicenters. Section I gives credit to seeds posted by authors or domains of organizations near New York City. Consequently, the top five seeds were posted by authors (e.g., @NYGovCuomo - Governor of New York and @seanhannity - talk show host and conservative political commentator) and domains of organizations (e.g., mediaite.com, nytimes.com, and newyorker.com) resident in New York. Section I also highlights stories about the pandemic in United States: *Dr. Fauci Shoots Down ‘Conspiracy Theory’ That Coronavirus Deaths Are Being Inflated: ‘No Evidence That’s the Case At All’* - mediaite.com, and *Will the Coronavirus Kill the Oil Industry?* - newyorker.com.

Section III of Table 28, just like Section I of the same table, gives credit to seeds posted by authors or domain of organizations near London. Similar to Section I (with respect to New York), the top five seeds were posted by authors (e.g., @BBCNews and @MattCartoonist - Cartoonist at the Telegraph) and domains of organizations (e.g, bbc.co.uk news.sky.com, and theguardian.com) resident in London. This section also highlights issues about the pandemic in the UK, such as: *Coronavirus: Boris Johnson moved out of intensive care but remains in hospital* - news.sky.com. Section II of Table 28 flips the reward system of Section I to prioritize authors and domains distant from New York City. This resulted in the surfacing of authors and domains outside the United States. The top five authors are residents of two different countries (e.g, @ick_forPH - Philippines and @OfficialKRU - Kenya) while the organization of the domains are from four different countries (thejakartapost.com - Indonesia, rappler.com - Philippines, bylinetimes.com - England, and kru.co.ke, jesusislordradio.info - Kenya). Changing the geographical focus from New York, in

TABLE 28: For *The 2020 Coronavirus Pandemic*, top five seeds extracted by combining geographical QPs ge_a and ge_d , ranking the seeds by their QP scores (q - Equation 13), and selecting the top five seeds with the highest scores. The table illustrates the interplay of ge_a and ge_d by showing how authors from different geographical regions share seeds from different domains.

#	domain (domain org. location): title (author's twitter handle, author's location)	QP Normalized			QP (Miles)	
		q	ge_a	ge_d	ge_a	ge_d
Section I (epicenter, New York City)						
1	ny.gov (New York): The Official Website of New York State (@NYGovCuomo, New York)	1.00	1.00	1.00	0.00	0.00
2	mediaite.com (NYC): Dr. Anthony Fauci Denies Claims Coronavirus Deaths Inflated (@oLiverdarcy, NYC)	1.00	1.00	1.00	0.00	0.00
3	nytimes.com (NYC): Hospitals Warn Nurses and Doctors Not to Speak Out on Coronavirus (@nytimes, NYC)	1.00	1.00	1.00	0.00	0.00
4	hannity.com (NYC): DEVELOPING: Dems Block McConnells \$250B Aid Package for Small Businesses During Coronavirus (@seanhannity, NYC)	1.00	1.00	1.00	0.00	0.00
5	newyorker.com (NYC): Will the Coronavirus Kill the Oil Industry? (@NewYorker, NYC)	1.00	1.00	1.00	0.00	0.00
Section II (epicenter, New York City)						
		q	$\overline{ge_a}$	$\overline{ge_d}$	$\overline{ge_a}$	$\overline{ge_d}$
1	thekartapost.com (Jakarta): Finland discovers masks bought from China not hospital-safe (@ick_forPH, Philippines)	0.92	0.83	1.00	8,598	10,051
2	rappler.com (Philippines): FACT CHECK: Duque claims PH has 'low' coronavirus infection(@rapplerdotcom, Philippines)	0.84	0.83	0.86	8,598	8,598
3	bylinetimes.com (London): COVID-19 SPECIAL INVESTIGATION: Leaked Home Office Call Reveals Government wants Economy to 'Continue Running' as 'We Will All Get' COVID-19 Anyway(@GHNeale, NA)	0.75	1.00	0.34	10,397	3,461
4	kru.co.ke (Nairobi): Kenya Rugby Union announces cancellation of 2019/20 season as Corona virus continues to hit sport (@OfficialKRU, Nairobi)	0.72	0.71	0.73	7,358	7,358
5	jesuslordradio.info (Nakuru, Kenya): Welcome To Jesus Is Lord Radio(@_lameckonger, Kisii, Kenya)	0.71	0.69	0.72	7,225	7,274
Section III (epicenter, London)						
		q	ge_a	ge_d	ge_a	ge_d
1	bbc.co.uk (London): Coronavirus: BBC presenter Emily Maitlis criticises 'misleading' language (@BBCNews, London)	1.00	1.00	1.00	0.00	0.00
2	news.sky.com (London): Coronavirus LIVE: UK claps to say thank you to NHS workers fighting coronavirus (@SkyNews, London)	1.00	1.00	1.00	0.00	0.00
3	theguardian.com (London): Coronavirus is the greatest global science policy failure in a generation (@Littlecub647, London)	1.00	1.00	1.00	0.00	0.00
4	telegraph.co.uk (London): Matt (@MattCartoonist, London)	1.00	1.00	1.00	0.00	0.00
5	news.sky.com (London): Coronavirus: Boris Johnson moved out of intensive care but remains in hospital (@SkyNews, London)	1.00	1.00	1.00	0.00	0.00



Fig. 42: The seed: <https://jesusislordradio.info/> (Table 28, Section II, No. 5) shared by @_lameckonger from Kisii, Kenya, was surfaced by prioritizing authors (\overline{ge}_a) and domains (\overline{ge}_d) distant from New York City. This seed could be considered non-relevant, however, if the context of concern requires supplying domains that satisfy the condition *religious responses to the Coronavirus Pandemic*, the seed could be considered relevant.

this case, shifted the perspective of the news away from the United States as Section II also reveals. Unlike Section I, which focused on the pandemic from the perspective of the United States, Section II shifts the focus to other countries, for example, *Finland discovers masks bought from China not hospital-safe* - thejakartapost.com (Section II, No. 1), *Kenya Rugby Union announces cancellation of 2019/20 season as Corona virus continues to hit sport* - kru.co.ke (Section II, No. 4). The non-relevant seed with the title *Welcome To Jesus Is Lord Radio* - jesusislordradio.info (Section II, No. 5) illustrates the effect

of not including relevance in the QP score q of seeds. However, relevance can be subjective, and thus, it is not inconceivable that this seed could be considered relevant (adding to the diversity of seeds) if the context (Figure 42) of concern is *religious responses to the Coronavirus Pandemic*. Jesus Is Lord Radio (jesusislordradio.info) was shared as the venue for a religious event about the pandemic.

TABLE 29: For *The 2020 Coronavirus Pandemic*, top five seeds with the highest *broad reputation* QP score (re_b). For a single seed (e.g., <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1592694/>), the re_b score (e.g., 0.81) was approximated by counting the number of times the seed domain (e.g., nih.gov) was cited (e.g., 46 times) in a reputation gold standard of 57 representative Wikipedia documents (one vote per document) about *Disease outbreaks*.

#	domain: title (author’s twitter handle)	QP Normalized	QP (Hits)
		re_b	re_b
1	who.int : Tobacco (@SergioBowers1)	0.82	47
2	nih.gov : Ventilator-Associated Pneumonia: Diagnosis, Treatment, and Prevention (@HITNTNotTalkin)	0.81	46
3	cdc.gov : 2009 H1N1 Pandemic (H1N1pdm09 virus) Pandemic Influenza (Flu) (@2020DoOver)	0.74	42
4	cdc.gov : Legal Authorities for Isolation and Quarantine (@peabodypress)	0.74	42
5	cdc.gov : 2019-2020 U.S. Flu Season: Preliminary Burden Estimates (@Rick51224214)	0.74	42

Given the concerns [185, 186, 187, 188, 189, 190] surrounding the spread of misinformation/disinformation surrounding the coronavirus pandemic, curators could potentially impose stringent rules that restrict the sources of seeds to reputable sources. This *reputable sources only* selection criteria aligns with the goal of the *broad reputation* QP (re_b). Table 29 outlines the top five seeds when seeds are scored by their respective reputation scores. For a single seed (e.g., <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1592694/>) in Table 29, the re_b score (e.g., 0.81) was approximated by counting the number of times the seed domain (e.g., nih.gov) was cited (e.g., 46 times) in a reputation gold standard of 57

representative Wikipedia documents (one vote per document) about *Disease outbreaks*. Accordingly, the most dominant seeds were from world-renowned health institutions such as the World Health Organization (`who.int`) which was referenced 47 times, National Institute of Health (`nih.gov`) referenced 46 times, and Centers for Disease Control and Prevention (`cdc.gov`) referenced 42 times out of 57 representative Wikipedia documents about public disease outbreaks.

8.2 THE FLINT WATER CRISIS

This section references Tables 30 and 31, which report on seeds sampled from a collection of 384 seeds extracted on 2018-07-20 from Micro-collections from the Twitter-Top vertical with the “flint water crisis” query. In Chapter 1.2, we discussed the importance of extracting seeds from local news media especially for local events by highlighting how local news media reported the Flint story from the beginning while the national media was late to reporting the story. Table 30, Section I illustrates how the ge_d Quality Proxy can help us surface local news media organizations, such as `m1ive.com`, which was critical to the coverage of the *Flint Water Crisis*, by giving credit to seed domains from organizations near a geographical reference (e.g., Flint, Michigan). In contrast, flipping the geographical QP ($\overline{ge_d}$) rewarded distant news organization and surfaced foreign news media (e.g., `bbc.com`, `theguardian.com`, `who.int`) that could provide an international perspective of the *Flint Water Crisis*.

From Table 29, we first saw how re_b helped surface reputable (widely referenced) sources (e.g., `who.int`, `nih.gov`, and `cdc.gov`) for the Coronavirus Pandemic. Table 31, Section I illustrates the result of applying the same broadly-defined reputation Quality Proxy to the *Flint Water Crisis* story. Unsurprisingly, `nih.gov`, and multiple national/international news media such as `nytimes.com`, `cnn.com`, `reuters.com` displaced local media since these organizations are widely referenced (broad reputation) in Wikipedia documents, which is exactly the signal that re_b measures. In this table, for re_b , $hits$ represents the count of Wikipedia documents (one document - one vote) that cite a domain from a gold standard collection of 70 *Public health crisis* Wikipedia reference document. For re_n , it represents the number of times (out of 550 references) a domain was cited in the Wikipedia *Flint Water Crisis* document. Given reports [31] that the national media was late to report the Flint story, a curator might decide to prioritize locally-reputable source, those with a narrowly-defined reputation (re_n), which measure how often a domain is referenced for a specific story. The effect of prioritizing locally-reputable source is illustrated by Table 31,

TABLE 30: For *The Flint Water Crisis*, top five seeds for different domains extracted by combining relevance rl and the unflipped (ge_d) and flipped ($\overline{ge_d}$) geographical QP. ge_d (Section I) helped in surfacing local media (e.g., `mlive.com` and `detroitnews.com`) while $\overline{ge_d}$ rewarded seeds from news media organizations distant (e.g., `bbc.com` and `theguardian.com`) from Flint, Michigan.

#	domain (domain org. location): title (author’s twitter handle, author’s location)	QP Normalized			QP (Miles)
		q	rl	ge_d	ge_d
Section I					
1	<code>mlive.com</code> (Michigan): As Flint was slowly poisoned, Snyder’s inner circle failed to act (@PhilRevard, Michigan)	0.91	0.85	0.97	130.52
2	<code>eclectablog.com</code> (Ann Arbor, Michigan): The deceptive corporatist rewriting of the history of the #FlintWaterCrisis is in full swing (@LOLGOP, Ann Arbor, Michigan)	0.86	0.72	0.99	53.77
3	<code>detroitnews.com</code> (Detroit, Michigan): AG’s office got Flint complaints a year before probe (@PhilRevard, Michigan)	0.85	0.68	0.99	57.96
4	<code>michiganadvance.com</code> (Michigan): Judge allows Flint water class-action lawsuit to proceed, adds Snyder back as defendant (@jmlarkin, Cambridge, MA)	0.84	0.70	0.97	130.52
5	<code>michigan.gov</code> (Michigan): EGLE - Flint’s water remains stable, continues to meet federal and new stricter state standards (@nreza21, NA)	0.84	0.69	0.97	130.52
Section II					
		q	rl	$\overline{ge_d}$	$\overline{ge_d}$
1	<code>bbc.com</code> (London): Flint water crisis: Prosecutors drop all criminal charges (@BBCWorld, London)	0.80	0.68	0.90	3,745
2	<code>theguardian.com</code> (London): The Flint water crisis is a shadow on Obama’s legacy (@Synthdrum, San Jose, CA)	0.78	0.64	0.90	3,745
3	<code>who.int</code> (Geneva): Lead in Drinking-water (@PeterMaier36, Stansbury, UT)	0.71	0.00	1.00	4,173
4	<code>wikipedia.org</code> (NA): Flint water crisis (@SuperSpacedad, NA)	0.64	0.91	NA	NA
5	<code>independent.co.uk</code> (London): Georgina Bloomberg: I’m grateful my dad didn’t run against Donald Trump (@brianpmangan, NYC)	0.64	0.06	0.90	3,745

Section II which gives credit to multiple local media organizations (`mlive.com`, `abc12.com`, `freep.com`, etc.) that covered the Flint story from its genesis.

8.3 HURRICANE HARVEY

This section references Tables 32 which sampled seeds from a collection of 384 seeds extracted on 2017-09-01 from tweets extracted from Twitter-Top Micro-collections with the “hurricane harvey” query.

The vast majority of important stories and events reported, are reported by news organization. So it comes as no surprise that the seeds extracted for various stories consists

TABLE 31: For *The Flint Water Crisis*, top five seeds extracted by focusing on broad (referenced across *Public health crisis* topics - Section I, re_b) and narrow (referenced only in the *Flint Water Crisis* story - Section II, re_n) reputation. For re_b , **Hits** represents the count of Wikipedia documents (one document - one vote) that cite a domain from a gold standard collection of 70 *Public health crisis* Wikipedia reference documents. For re_n , it represents the number of times (out of 550 references) a domain was cited in the Wikipedia *Flint Water Crisis* document. Broadly-defined reputation benefits well-known (e.g., `nih.gov` and `nytimes.com`) organizations. Narrowly-defined reputation benefits local media (e.g., `mlive.com`, `abc12.com`, `freep.com`).

#	domain (author's twitter handle): title	QP Normalized	QP (Hits)
Section I		re_b	re_b
1	<code>nih.gov</code> : Flint Water Crisis: What Happened and Why? (@SaigeTucker)	0.39	27
2	<code>nytimes.com</code> : Flint Water Prosecutors Drop Criminal Charges, With Plans to Keep Investigating (@FireGoddessB)	0.37	26
3	<code>cnn.com</code> : Miss Michigan calls out Flint water issue (@CNN)	0.33	23
4	<code>reuters.com</code> : The thousands of U.S. locales where lead poisoning is worse than in Flint (@dfi_playah)	0.29	20
5	<code>abcnews.go.com</code> : Lapses at all levels of government made Flint water crisis worse: Watchdog (@RedTRaccoon)	0.26	18
Section II		re_n	re_n
1	<code>mlive.com</code> : As Flint was slowly poisoned, Snyder's inner circle failed to act (@PhilRevard)	0.41	226
2	<code>abc12.com</code> : Flint gets \$77 million to pay for water projects (@DavidLeonMorgan)	0.08	42
3	<code>freep.com</code> : All Flint's children must be treated as exposed to lead (@lovetogive2)	0.06	32
4	<code>detroitnews.com</code> : AG's office got Flint complaints a year before probe (@PhilRevard)	0.04	22
5	<code>nih.gov</code> : Flint Water Crisis: What Happened and Why? (@SaigeTucker)	0.03	18

mainly of seeds from well-known news media organizations (e.g., *CNN* and *FoxNews*). However, far-reaching new stories such as the catastrophic *Hurricane Harvey of 2017* are also covered by media outlets (e.g., *eonline.com* and *espn.com*) with a different focus, such as Sports or Entertainment, offering a different perspective from conventional news media. The scarcity *sc* Quality Proxy attempts to identify such media outlets based on the premise that their seeds are scarce. Table 32 (Section I) illustrates the application of *sc* to surface seeds from non-conventional news media outlets such as *Taylor Swift Makes “Very Sizable Donation” to Houston Food Bank After Hurricane Harvey* - *eonline.com* and *J.J. Watt’s Hurricane Harvey charity fundraising closes with \$37M-plus in donations* - *espn.com*. Section I of Table 32 contrasts the entertainment sources (e.g., *eonline.com*, *espn.com*, and *rollingstone.com*) with more conventional news sources (e.g., *cnn.com*, *abcnews.go.com*, and *washingtonpost.com*) found in Section II by flipping (\overline{sc}) the scarcity QP.

8.4 CHAPTER SUMMARY

In this chapter, we explored how different combinations of Quality Proxies interact to surface seeds of different characteristics. This was achieved by extracting seeds from Micro-collection in Twitter for *The 2020 Coronavirus Pandemic*, *The Flint Water Crisis*, and *Hurricane Harvey*. Next, we assigned Quality Proxy scores to the seeds with different combination of Quality Proxies ($\{rp, lk, sh\}$, $\{ge_a, ge_d\}$, $\{\overline{ge_a}, \overline{ge_d}\}$, $\{re_n\}$, $\{re_b\}$, etc). We showed that seeds selected by different combinations of QP scores map to different policies (e.g., prioritizing popularity - $\{rp, lk, sh\}$ or narrow reputation - $\{re_n\}$). Different QP score combinations fulfill different seed selection goals, illustrating the versatility in seed selection the Quality Proxies offer.

TABLE 32: For *Hurricane Harvey*, top five seeds extracted by combining relevance (rl) and the unflipped scarcity Quality Proxy sc (Section I) and flipped \overline{sc} (Section II). Scarcity can be used to increase the diversity of the seed domains as reflected by the domains (e.g, texasmonthly.com, eonline.com, and espn.com), flipping the Quality Proxy results in surfacing seeds from domains (e.g., cnn.com and abcnews.go.com) that appear multiple times in the collection.

#	domain: title (author’s twitter handle, author’s location)	QP Normalized			QP (Hits)
		q	rl	sc	sc
Section I					
1	texasmonthly.com : Voices from the Storm (@TexasMonthly)	0.71	0.13	0.99	1
2	texasobserver.org : Even Hurricane Harvey Can’t Temper GOP Hostility Toward Texas’ Big Cities (@texasdemocrats)	0.70	0.11	0.99	1
3	eonline.com : Taylor Swift Makes “Very Sizable Donation” to Houston Food Bank After Hurricane Harvey (@enews)	0.70	0.10	0.99	1
4	espn.com : J.J. Watt’s Hurricane Harvey charity fundraising closes with \$37M-plus in donations (@SportsCenter)	0.70	0.08	0.99	1
5	rollingstone.com : Houston Astros After Hurricane Harvey (@RollingStone)	0.70	0.08	0.99	1
Section II					
		q	rl	\overline{sc}	\overline{sc}
1	cnn.com : Harvey aftermath: Toxic waste sites flooded in Texas, EPA says (@cnnbrk)	0.11	0.15	0.03	3
2	abcnews.go.com : Hurricane Harvey wreaks historic devastation: By the numbers (@ABC)	0.10	0.14	0.05	5
3	texasmonthly.com : Voices from the Storm (@TexasMonthly)	0.09	0.13	0.01	1
4	washingtonpost.com : Displaced pets from Hurricane Harvey and now Irma need our help (@washingtonpost)	0.08	0.12	0.02	2
5	climaterealityproject.org : Hurricane Harvey and Climate Change: Here Are the Facts (@CouchCoopParent)	0.08	0.11	0.02	2

CHAPTER 9

A FRAMEWORK FOR BOOTSTRAPPING WEB ARCHIVE COLLECTIONS FROM MICRO-COLLECTIONS IN SOCIAL MEDIA

The contributions of the previous chapters provide the foundation for this chapter in which we present our framework for bootstrapping Web archive collections from Micro-collections in social media. In Chapter 5 we introduced social media Micro-collections as a novel source for seeds extracted from the threaded conversations of single or multiple authors. We also showed that Micro-collections express editorial activity that may be indicative of their quality. In Chapter 6 we took the first step toward providing a generic approach for profiling and comparing collections of seeds by proposing the Collection Characterizing Suite (CCS). In Chapter 7 we advanced the contributions of Chapter 6 from collection comparison to quality evaluation through the introduction of the Quality Proxies (QP) for seeds. In Chapter 8 we showed how different combinations (e.g., $\{rp, sh, lk\}$, $\{geo_a, geo_d\}$, $\{\overline{geo_a}, \overline{geo_d}\}$) of Quality Proxies map to different policies for selecting seeds, and thus, surface seeds corresponding to the semantics of the QP combination.

In this chapter, we build upon these contributions and present our Micro-collection/Quality Proxy (MCQP) framework for bootstrapping Web archive collections from Micro-collections in social media. Next, having already assessed the other components of the framework in Chapters 5 and 6, we focus on evaluating the seeds selected using scores assigned by Quality Proxies.

9.1 FRAMEWORK OVERVIEW

As Figure 43, the framework overview illustrates, the seed generation process for bootstrapping Web archive collections begins with the query. All “Stage” references in this section refer to Figure 43. A query representing the story or event is issued (Stage 1) to social media SERPs such as Reddit, Twitter, Facebook, and a placeholder for expansion to additional social media platforms as they become available. The placeholder is important because social media platforms come and go. For example, in the middle of this work,

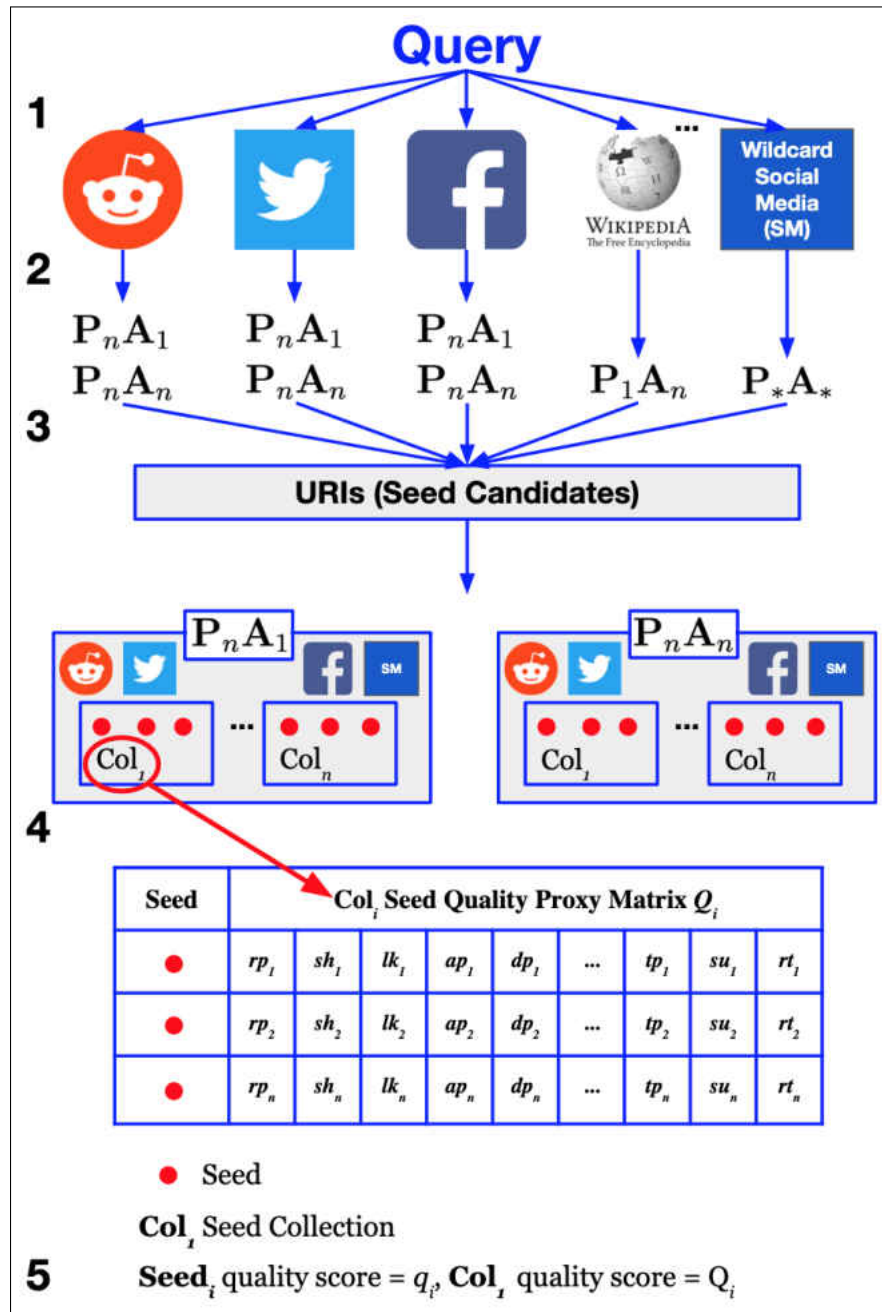


Fig. 43: MCQP framework overview for bootstrapping Web archive collections from Micro-collections in Social Media. The numbers shown represent the stages of the framework.

Storify went out of service [36], resulting in the use of Scoop.it as a substitute for our study in Chapter 5. URIs from the social media posts from Stage 1 are extracted from social media posts with replies authored by a single ($P_n A_1$) or multiple ($P_n A_n$) authors (Figure

43, Stage 2). If a canonical Wikipedia page for the story exists, URIs from the references section of the Wikipedia page (which belong to the $\mathbf{P}_1\mathbf{A}_n$ post class) may be extracted as seed candidates. All the URIs extracted make up the seed candidates list (Stage 3).

$$q = |\mathbf{q}|_2 = \sqrt{\sum_{i=1}^n q_i^2} \quad (15)$$

Repetition of Equation 13

In Stage 4, the quality of each seed is determined by generating the seed QP vector (Chapter 7) for the seed. The n -dimensional seed Quality Proxy vector \mathbf{q} of a seed $seed_i$ from a Micro-collection col_i expresses a quality trait of a seed across each dimension. A group of Quality Proxy vectors make up the Quality Proxy matrix \mathbf{Q} which expresses the quality of all the seeds. The Quality Proxy score (Equation 15, Stage 5) of the seed q or collections Q is a cumulative expression of the quality of the seed or the collection. As we saw in the previous chapter, different combinations of QPs map to different semantics and policies of selecting seeds. Consequently, the final stage of the framework provides the ability of the user to utilize a given combination of QPs to assign quality scores to seeds, and subsequently select the top K seeds with the highest scores. This is how we evaluated the Quality Proxies, by assessing the quality of the top K seeds selected by different combinations of QPs.

9.2 FRAMEWORK EVALUATION

The goal of this evaluation was two-fold. The first goal was to assess the precision of the seeds selected by Quality Proxies when novelty is not prioritized (Chapter 9.2.1). Since we propose the use of seeds extracted from Micro-collections to bootstrap collection building or augment expert-generated seeds, it is crucial to assess the quality of these seeds. It would be unreasonable to use seeds from Micro-collections, selected based on their Quality Proxies scores, if they are of poor quality compared to expert-generated seeds. Good quality was modeled by prototypical seeds referred to as *reference* seeds selected from Google or hand-selected by human-experts on Archive-It.

The second goal of the evaluation was to assess the precision of seeds when novelty is prioritized (Chapter 9.2.2). It is a positive trait for seeds selected by QPs from Micro-collections to be highly similar (low novelty) with respect to Google and/or expert-generated seeds (Table 33), since this could be indicative of their high-quality. However, we often need seeds from Micro-collections to be novel or, in other words, different from seeds produced by Google and/or experts. Nevertheless, quality must not be compromised for novelty. Therefore, the second goal of the evaluation was to assess the precision of seeds selected by

QPs from Micro-collections when novelty is prioritized. Novelty of Micro-collection seeds was measured by comparing them with reference (Google or Expert) seeds. We extended the idea of measuring novelty by quantifying the diversity or variety (Chapter 9.2.3) of seeds for independent collections (reference and Micro-collections) selected with QPs.

To evaluate seeds selected by Quality Proxies from Micro-collections, we generated a dataset (Table 33, [191]) consisting of seeds extracted from reference collections and Micro-collections for multiple topics. The reference collections from Google (*All* vertical) and expert-generated collections (from Archive-It) served as baselines for defining quality. Additionally, novelty was measured with respect to these reference collections. Seeds from Google and Twitter were scraped, while seeds from expert-generated collections were extracted from the Archive-It API [192]. As outlined by the **Extraction-Range** field of Table 33, some seeds were collected at the same time, while others were collected periodically over the specified date range. In total, the dataset consisted of 1,552 seeds from Reference (Google and Expert) collections, and 2,027 seeds from 4,209 tweets from Twitter Top/Latest Micro-collections extracted at different date ranges.

9.2.1 STEPS FOR ASSESSING SEED PRECISION WHEN NOVELTY IS NOT PRIORITIZED

The following five steps describe how we assessed the precision of seeds without prioritizing novelty.

Step 1: Extracting Quality Proxies for Seeds

For all seeds in the evaluation dataset we extracted (Chapter 7) 12 QP measures (Table 34): *reply* (rp), *share* (sh), *like* (lk), *author-popularity* (ap), *domain-popularity* (dp), *geographical-author* (ge_a), *geographical-domain* (ge_d), *retrievability* (rt), *scarcity* (sc), *reputation-broad* (re_b), *reputation-narrow* (re_n), and *relevance* (rl). The *subject-expert* (su) QP instantiation with the document frequency from Google was not determined to be a dependable approximation of su since it fluctuated with a high variance, hence we excluded it from our evaluation. Additionally, we chose not to impose a temporal bias to favor old or new documents, hence we excluded the *temporal* (tp) QP.

We approximated the *relevance* QP with similarity between a seed’s document vector and a gold standard document vector created from the text extracted from the references of Wikipedia articles (Table 35) corresponding to each dataset topic. The *author-popularity* ap QP corresponds to the popularity of the social media author of the post. Since seeds

TABLE 33: Framework evaluation dataset [191] consisting of 1,552 seeds from Reference (Google & Expert) collections, and 2,027 seeds from 4,209 tweets from Twitter Top/Latest Micro-collections extracted at different date ranges.

Topic	Extraction-Range	Curator	Seeds Count
Reference Google Collections (808 Seeds)			
hurricane harvey	2020-04-11	Nwala	199 (Pages 1 - 20)
flint water crisis	2020-04-10	Nwala	173 (Pages 1 - 20)
coronavirus	2020-04-09	Nwala	176 (Pages 1 - 20)
2018 world cup	2019-01-09	Nwala	112 (Pages 1 - 10)
ebola virus	2017-11-29	Nwala	97 (Pages 1 - 10)
hurricane harvey	2017-09-02 to 2017-09-29	Nwala	51 (Page 1)
Reference Expert Collection from Archive-It (744 Seeds)			
coronavirus [193]	2020-03-15	NLM	574
hurricane harvey [194]	2017-08-25 to 2017-09-29	VTech	37
ebola virus [195]	2014-10-01	NLM	133
Micro-collections from Twitter-Top (1,310 Seeds, 2,221 tweets)			
hurricane harvey	2020-04-11	Nwala	201 (500 tweets)
flint water crisis	2020-04-09	Nwala	312 (500 tweets)
coronavirus	2020-04-09	Nwala	533 (500 tweets)
2018 world cup	2019-01-09	Nwala	121 (500 tweets)
ebola virus	2017-11-30 to 2017-12-31	Nwala	48 (68 tweets)
hurricane harvey	2017-09-02 to 2017-09-31	Nwala	95 (153 tweets)
Micro-collections from Twitter-Latest (717 Seeds, 1,988 tweets)			
flint water crisis	2020-04-09	Nwala	92 (500 tweet)
coronavirus	2020-04-09	Nwala	541 (500 tweet)
2018 world cup	2019-01-09	Nwala	84 (488 tweet)

TABLE 34: List of Quality Proxies extracted from evaluation dataset seeds. We additionally included the flipped states of these Quality Proxies for scoring seeds.

#	Quality Proxies
1	<i>reply (rp)</i>
2	<i>share (sh)</i>
3	<i>like (lk)</i>
4	<i>author-popularity (ap)</i>
5	<i>domain-popularity (dp)</i>
6	<i>geographical-author (ge_a)</i>
7	<i>geographical-domain (ge_d)</i>
8	<i>retrievability (rt)</i>
9	<i>scarcity (sc)</i>
10	<i>reputation-broad (re_b)</i>
11	<i>reputation-narrow (re_n)</i>
12	<i>relevance (rl)</i>

from Google are not posted by social media authors, we approximated the *ap* QP with the reciprocal rank ($\frac{1}{rank_i}$) of the seed. Similarly, for expert-generated seeds, we approximated *ap* with the reciprocal position (or rank) of the seed in the collection.

Step 2: Generating Quality Proxies Combinatorial States

The reference (Google and Expert) collections serve as quality baselines for comparing Micro-collections. Given different groups of seeds, we can check if the seeds extracted from Micro-collections are similar (e.g., from similar domains) to those from Google or Archive-It. However, such method of comparison gives no room for assessing the Quality Proxies since all the seeds from the Micro-collections are selected. In contrast, given K top (e.g., “top” defined according to their rank in the SERPs or QP scores) seeds from Google, we can check if the top K seeds from Micro-collections are similar (e.g., from similar domains) to those from Google. The task of selecting the top K seeds from Micro-collections requires a means of scoring the seeds which we achieved by utilizing QPs (Table 34).

We utilized the 12 QPs from Table 34 to score (Equation 15) seeds, selected the top K seeds, and compared them with top reference seeds scored with the same QPs. We did not

TABLE 35: Precision gold-standard dataset. The documents from the references of these Wikipedia articles were used to generate document vectors for measuring relevance. Relevance was approximated by the similarity between a seed’s document vector and the gold-standard vector corresponding to the seed’s topic. Similarity exceeding the specified relevance threshold signaled the relevance of the seed.

Topic	Wikipedia Reference	Relevance Threshold	Document Count
hurricane harvey	https://en.wikipedia.org/wiki/Hurricane_Harvey	0.10	183
flint water crisis	https://en.wikipedia.org/wiki/Flint_water_crisis	0.20	550
coronavirus	https://en.wikipedia.org/wiki/COVID-19_pandemic	0.20	719
2018 world cup	https://en.wikipedia.org/wiki/2018_FIFA_World_Cup	0.20	400
ebola virus	https://en.wikipedia.org/wiki/Western.African.Ebola.virus.epidemic	0.20	697

assign weights (Chapter 7.5) to the Quality Proxies. Additionally, we expanded the options for scoring seeds beyond 12 QPs as follows. First, we permitted flipping the QPs, resulting in 12 additional QPs (24 QPs total). But it should be noted that an unflipped QP (e.g., ge_a) cannot be combined with its flipped version ($\overline{ge_a}$), because these states are mutually exclusive. Second, we permitted using a subset of the 24 QPs, leading to a combinatorial explosion of possible QP states for scoring seeds. However, we restricted our scoring to 1-, 2-, and 3-combinations which produced a total of 2,049 possible QP combinations. Table 36 shows a small sample of the different r -combinations ($r \in [1, 3]$) which present different ways of selecting the QPs to score (Equation 15) seeds.

Step 3: Scoring Seeds with a Combination of Quality Proxies

To score seeds from Micro-collections or reference Google or Expert collections, first, we first selected a single combination of Quality Proxies, for example, rp, lk, sh . Next, using only the QPs selected, we assigned a score to the seed with Equation 15.

Step 4: Top K seeds comparison: Micro-collections vs reference collections

We sorted all the seeds extracted from Micro-collections in descending order of their scores. Next, the top K seeds, ranked by their respective QP scores assigned by a given combination of QPs, were compared to the top K seeds of reference collections processed in

TABLE 36: A sample of 12 QP combinatorial states for 1-combination, 2-combination, and 3-combinations. A single 1-combination or 2-combination or r -combination of QPs can be used to score (Equation 15) a seed.

#	1 - Combination	2 - Combinations	3 - Combinations
1	rp	rp, lk	rp, lk, sh
2	$\overline{r\overline{p}}$	rp, sh	rp, lk, ap
3	sh	rp, ap	rp, lk, dp
4	$\overline{s\overline{h}}$	rp, dp	rp, lk, ge_a
5	lk	rp, ge_a	rp, lk, ge_d
6	$\overline{l\overline{k}}$	rp, ge_d	rp, sh, ap
7	ap	lk, sh	rp, sh, dp
8	$\overline{a\overline{p}}$	lk, ap	rp, sh, ge_a
9	dp	lk, dp	rp, sh, ge_d
10	$\overline{d\overline{p}}$	lk, ge_a	rp, ap, dp
11	ge_a	lk, ge_d	rp, ap, ge_a
12	$\overline{g\overline{e}_a}$	sh, ap	rp, ap, ge_d

the same fashion. Comparison was done by measuring the domain overlap ($\frac{|A \cap B|}{\min(|A|, |B|)}$) between Micro-collection seeds and reference (Google and/or expert) seeds and their precision ($\frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{retrieved documents}|}$). For precision evaluation, if the similarity between a seed and the gold standard document vector is at least a predefined relevance threshold (Table 35), the seed is considered relevant. The threshold was estimated by finding the median similarity between each gold standard document and the rest of the gold standard documents. Median scores exceeding 0.20 — which was empirically determined to produce satisfactory baseline relevance — were set to 0.20.

Step 5: Assessing Seed Precision when Novelty is not Prioritized

The final process of assessing the precision of seeds when novelty is not prioritized involved reporting the overlap and precision for QP combinations used to score (and select top K) seeds. This was achieved by reporting the top 10 overlap scores between Micro-collection and reference seeds and reporting Precision at K (P@K) for the associated QP combination used to score the seeds. Selecting the top 10 overlap enables us learn the precision of seeds

when overlap is at its best, albeit at the expense of novelty since the higher the overlap between Micro-collection and reference seeds, the lower the novelty. Chapter 9.3.1 presents and discusses the results for assessing seed precision when novelty is not prioritized.

9.2.2 STEPS FOR ASSESSING SEED PRECISION WHEN NOVELTY IS PRIORITIZED

Since we consider reference seeds to be quality seeds, a high overlap between reference and Micro-collection seeds could result in a high precision of Micro-collection seeds. However, since novelty (low overlap) is also a desirable quality of seeds, it is crucial to additionally assess the precision of Micro-collection seeds (selected by QPs) when novelty is prioritized.

The steps for assessing the precision of seeds when novelty is prioritized is the same as the previous section (when novelty is not prioritized) except for the last step (*Step 5: Assessing Seed Precision when Novelty is not Prioritized*). Instead of reporting the P@K for the associated QP combinations with the top 10 overlap scores, to prioritize novelty, we measured and reported the precision of QP combinations that produced a low overlap (high novelty) between Micro-collection and reference seeds. Chapter 9.3.2 presents and discusses the results for assessing seed precision when novelty is prioritized.

9.2.3 ASSESSING DIVERSITY OF SEEDS FROM MICRO-COLLECTIONS

In an attempt to represent multiple views while collecting seeds, it is often a desirable attribute for the seeds collected to be diverse. For example, a collection about the 2018 *Kavanaugh hearings* — a highly partisan political story — that sampled seeds exclusively from `cnn.com` or exclusively from `foxnews.com` could be reasonably labeled to be skewed to the left (for `cnn.com` seeds) or right (for `foxnews.com`). While ensuring a collection is diverse by sampling from multiple domains (e.g., `cnn.com`, `foxnews.com`) does not guarantee a balancing of the multiple viewpoints of a story, it is a crude attempt to do so.

The different combinations of Quality Proxies represent different policies for selecting seeds as we saw in the previous chapter. Consequently, it is expected that the different combinations of QPs would result in different diversity scores for the seeds. To get a complete view of the distribution of diversity across all the Quality Proxies used to score seeds, we took the following steps.

First, for each of the evaluation dataset topics, we generated 2,049 1-, 2-, and 3-combinations of Quality Proxies. Table 36 shows a small fraction of the combinations. Second, given a single QP combination (e.g., *rp, ap, ge_d*) we used it to score (Equation 15)

reference (Google - G, Expert - E) and Micro-collection seeds (M_G and M_E ¹).

Third, we sorted the seeds in descending order with their respective QP scores, selected the top $K = 10$ seeds, and measured the diversity resulting from the utilization of the QP combination to score and subsequently select seeds. The smaller the value of K , the larger the diversity since the collections have a better chance at filling a fewer number of slots with unique entries. However, since the purpose of our evaluation was to measure the diversity of seeds from one collection (G or E) relative others (M), the value of K need only be the same across all collections.

Fourth, we repeated the previous step 2,049 times for all the 1 – 3 QP combinations to get the diversity of the top 10 seeds selected for each QP combination. The diversity d_u (introduced in Chapter 6.1.5) of a collection $|C| > 1$ of seeds is simply the ratio of unique seeds U to the total number of seeds in the collection: $\frac{U}{|C|}$. In addition to d_u , we measured diversity (d_c) by measuring the change in size of K seeds after compression using the LZW data compression algorithm. The rationale for how compression can be used to quantify diversity is explained by the following example. Consider two different collections of seeds, the first has half of K seeds from the same domain (e.g., `cnn.com`). In contrast, for the second collection, all K seeds are from distinct domains. After compression, the change in sizes of the first collection is more drastic than the second due to the compression of the repeated domain strings. Consequently, d_c is defined as follows where $|C_0|$ represents the size (string length) of a collection before it was compressed, while $|C_1|$ represents the size of the collection after compression: $1 - \frac{|C_0| - |C_1|}{|C_0|}$.

Alam et al. [196] introduced multiple URI normalization schemes that represent URIs by domain and path segment keys while stripping query parameters and fragments. We measured diversity at the domain level (paths excluded) even though other longer URI segments are possible. In addition to utilizing QP combinations to score seeds from G, E, M_G , and M_E , we randomly selected 10 seeds from each collection without using QP scores to discern the diversity resulting from not using QP scores. These collections have the r -superscript. Finally, we generated the Empirical Cumulative Distribution (Tables 44 and 45 and Appendices F and G) of the diversity of the QP combinations. The results are discussed in Chapter 9.3.4.

9.3 EVALUATION RESULTS AND DISCUSSION

The notation used in the tables and figures in this section to represent collections of

¹ M_G - represents Micro-collection seeds when the reference seeds are from Google, M_E - Expert reference.

TABLE 37: (Chapter 9.3.1, Coronavirus): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP Combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap				P@10								QP Combinations	
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$	GM	EM
1	1.0	1.0	.23	.27	.80	.80	.40	.63	.60	1.0	.83	.38	rl, re_b	re_n
2	1.0	1.0	.23	.20	.60	.80	.56	.44	.56	1.0	.43	.83	rt, re_b	rl, dp
3	1.0	1.0	.20	.20	.90	.80	.86	.88	.60	.80	.50	.43	rl, \overline{rp}, re_b	\overline{rl}, dp
4	1.0	1.0	.20	.19	.90	.80	.71	.86	.67	.90	1.0	.71	rl, \overline{sh}, re_b	\overline{ap}, dp
5	1.0	1.0	.20	.19	.90	.80	.89	.75	.67	.90	.40	.57	rl, \overline{lk}, re_b	dp, ge_a
6	1.0	1.0	.19	.19	.80	.80	.57	.88	.56	1.0	.38	.83	rl, \overline{ap}, re_b	\overline{rt}, re_n
7	1.0	1.0	.19	.18	.80	.80	.88	.44	.60	1.0	.71	.22	rl, rt, re_b	rl, \overline{ap}, dp
8	1.0	1.0	.19	.18	.70	.80	.78	.78	.60	.90	.33	.83	rl, \overline{rt}, re_b	rl, dp, ge_a
9	1.0	1.0	.19	.18	.80	.80	.60	.89	.60	1.0	.44	.71	rl, sc, re_b	rl, dp, \overline{ge}_a
10	1.0	1.0	.19	.18	.80	.80	.56	.75	.60	1.0	.63	.63	rl, \overline{sc}, re_b	rl, dp, \overline{ge}_d
Averages														
	1.0	1.0	.20	.19	.80	.80	.68	.73	.61	.95	.57	.61		

TABLE 38: Top 10 seeds from Micro-collections extracted by two different QP combinations: rl, re_b (left) and $rl, \overline{lk}, \overline{ge}_a, rt$ (right). The QP scores prefix the domains. The left with 1.0 overlap with Google (precision: 0.6) consists of popular domains (e.g., `nytimes.com` and `who.int`) while the right (0.0 overlap, 0.6 precision) consists of less popular and international (due to \overline{ge}_a) domains (e.g, `bylinetimes.com` and `rappler.com`). Non-relevant seeds have been struck through.

#	QP Combination: rl, re_b Overlap: 1.0, Precision: 0.6	QP Combination: $rl, \overline{lk}, \overline{ge}_a, rt$ Overlap: 0.0, Precision: 0.6
1	(0.609) <code>nytimes.com</code> : As New Coronavirus Spread, China’s Old Habits Delayed Fight (<code>@epallred, NA</code>)	(0.726) <code>bylinetimes.com</code> : COVID-19 SPECIAL INVESTIGATION: Leaked Home... (<code>@GHNeale, Kent, UK</code>)
2	(0.592) <code>who.int</code>: Tobacco (<code>@SergioBowers1, Alberta, Canada</code>)	(0.674) <code>rappler.com</code> : FACT CHECK: Duque claims PH has ‘low’ coronavirus infection (<code>@rapplerdotcom, Philippines</code>)
3	(0.584) <code>nytimes.com</code> : E.U. Officials Agree to Deal to Soften Coronaviruss Economic Blow (<code>@nytimes, New York</code>)	(0.661) <code>thejakartapost.com</code> : Finland discovers masks bought from China not hospital-safe (<code>@icklato, Philippines</code>)
4	(0.576) <code>nih.gov</code>: Ventilator-Associated Pneumonia: Diagnosis, Treatment, and Prevention (<code>@HITNTNotTalkin, NA</code>)	(0.657) <code>bbc.com</code> : The Chinese doctor who tried to warn others about coronavirus (<code>@PramodSpeaks, Mumbai, India</code>)
5	(0.566) <code>nytimes.com</code> : How Delays and Unheeded Warnings Hindered New Yorks Virus Fight (<code>@bpanz, S. California</code>)	(0.653) <code>theguardian.com</code> : UK coronavirus peak at least two weeks away, chief scientist says (<code>@SaldanhaWinston, NA</code>)
6	(0.546) <code>cdc.gov</code> : 2009 H1N1 Pandemic (H1N1pdm09 virus) (<code>@2020DoOver, US</code>)	(0.649) <code>independent.co.uk</code> : Coronavirus: More than 60 doctors have died during Italys outbreak (<code>@SaldanhaWinston, NA</code>)
7	(0.541) <code>nytimes.com</code> : Most New York Coronavirus Cases Came From Europe, Genomes Show (<code>@LincolnsBible, NA</code>)	<code>fiverr.com</code>: Draw your awesome cartoon portrait from your photo (<code>@Mubashsira2, Bangladesh</code>)
8	(0.539) <code>cdc.gov</code>: Legal Authorities for Isolation and Quarantine (<code>@peabodypress, Tulsa, OK</code>)	(0.628) <code>fiverr.com</code>: Cut out background removal clipping path (<code>@HaqBenjamin, Bangladesh</code>)
9	(0.529) <code>cdc.gov</code>: 2019-2020 U.S. Flu Season: Preliminary Burden Estimates (<code>@Rick51224214, NA</code>)	(0.628) <code>fiverr.com</code>: Do minimalist logo design to promote your business (<code>@Mubashsira2, Bangladesh</code>)
10	(0.525) <code>nytimes.com</code> : More Coronavirus Vaccines & Treatments Move Toward Human Trials (<code>@Curatorous, D.C., US</code>)	(0.628) <code>fiverr.com</code>: Access to This Page Has Been Blocked (<code>@RabbiKh94520371, Bangladesh</code>)

seeds, overlap, precision, and diversity are described as follows. The character G represents seeds generated from Google, E represents seeds collected by experts on Archive-It, and M represents seeds generated from Micro-collections in Twitter. These Micro-collections were extracted from the Twitter-Top vertical except when otherwise stated. The overlap between Google and Micro-collections is represented by GM, and Expert and Micro-collections by EM. Similarly, GM and EM also represent the combination of Quality Proxies used to score Google/Micro-collection and Expert/Micro-collection seeds, respectively. The overlap between random seeds from Google and Micro-collections is represented by $G^r M^r$, and random expert and Micro-collection seeds by $E^r M^r$. The overlap between Google or Expert (G or E) and Micro-collection (M) was calculated after scoring the seeds with Quality Proxies. In contrast, $G^r M^r$ and $E^r M^r$ represents the overlap without using Quality Proxies scores, to help facilitate estimating the improvement or deterioration of precision resulting from the utilization of QP scores.

The P@K of random (QP scores not used) seeds from Google is represented by G^r , while E^r represents the P@K of random Expert seeds. M_G represents the P@K of Micro-collection seeds when the reference collection is Google, while M_E represents the P@K of Micro-collection seeds when the reference collection is an Expert collection. $M^r G^r$ represents the P@K of random seeds from Micro-collections when the reference collections are random Google seeds, and $M^r E^r$ when the reference collections are random Expert seeds.

Our overlap and precision results were proven to be statistically significant by a one-tailed Student's t-test with $\alpha = 0.05$ and $K = 30$ across all dataset topics.

9.3.1 RESULTS: ASSESSING SEED PRECISION WHEN NOVELTY IS NOT PRIORITIZED

Table 37 (for *2020 Coronavirus Pandemic*) shows the top 10 overlap values between Google (G) and Micro-collection (M) seeds (GM), as well as the overlap between Expert (E) seeds and Micro-collection (GE) seeds. The same table shows the combination of Quality Proxies used to score the seeds. The GM field represents the QP combination used to score Google and Micro-collection seeds, while EM represents the QP combination for scoring Expert and Micro-collection seeds. In addition to the overlap, the table also shows the P@10 (P@10) for the top 10 seeds selected by their QP scores. There are five additional variants of Table 37 in Appendix A for the four remaining evaluation dataset topics. To help improve readability, the caption of these table are prefixed by information (red text) pointing to the section the table is meant for (e.g., Chapter 9.3.1) and the topic of the seeds

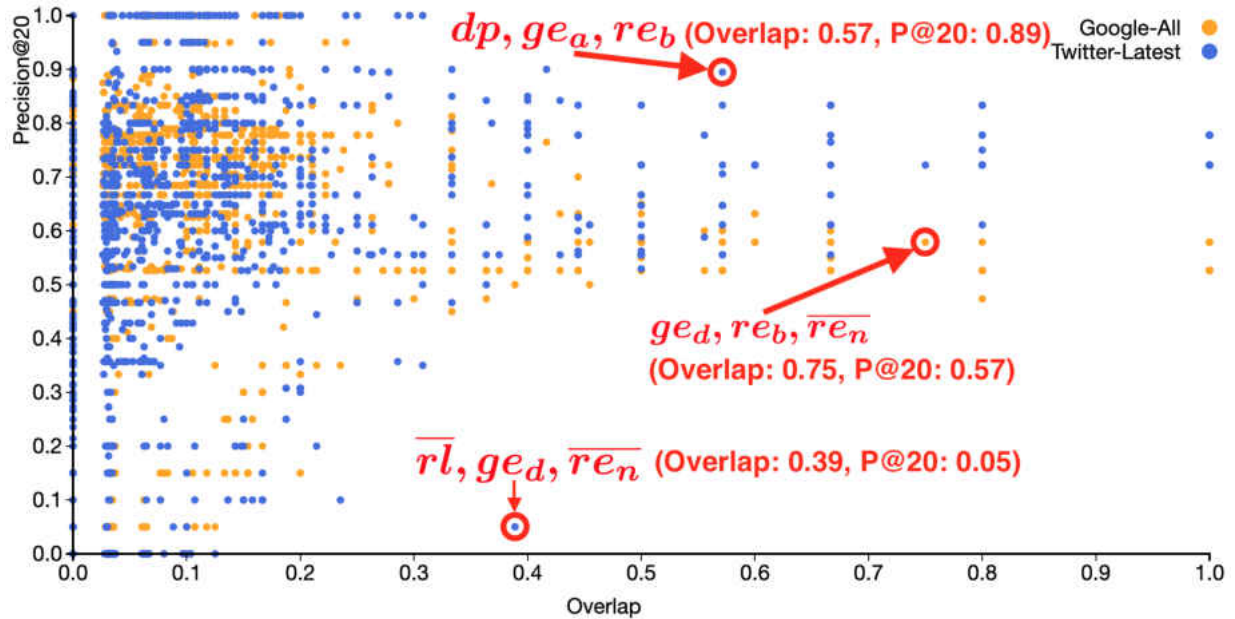


Fig. 44: Overlap vs P@20 for Google (G - orange dots) and Micro-collection (M - blue dots) *2020 Coronavirus Pandemic* Twitter-Latest seeds scored by different Quality Proxies. A single dot represents the overlap (X-axis) and P@20 (Y-axis) for seeds scored by a single Quality Proxy. The scatterplot shows how different Quality Proxy scores result in high (e.g., dp, ge_a, re_b and $ge_d, re_b, \overline{re_n}$) or low ($\overline{rl}, ge_d, \overline{re_n}$) overlap/P@20. Unsurprisingly, the QP combination $\overline{rl}, ge_d, \overline{re_n}$ resulted in a low P@20 because the *relevance rl* QP was flipped, meaning relevance was penalized.

(e.g., *2020 Coronavirus Pandemic*).

Overlap and P@10 for Google and Micro-collection seeds

Across all topics, for Google and Micro-collection seeds, the Minimum, Median, and Maximum (MMM) average overlap² were 0.24, 0.55, and 1.0, respectively, when Quality Proxies were used to score seeds. Without the utilization of QP scores, the MMM average overlap were smaller, 0.14, 0.17, and 0.27, respectively. These results from Table 37 and Appendix A suggest that the utilization of QP scores helped surface seeds from a common set of domains for Micro-collections and Google. The maximum overlap occurred for seeds from the *2020 Coronavirus Pandemic* topic (Table 37). This was not surprising since the seeds from Google and Micro-collections were authored and collected during a short period of time coinciding with the Coronavirus Pandemic. Table 38 (left column) shows a sample of 10 seeds scored by the QP combination rl, re_b that resulted in the overlap of 1.0³ caused by a common set of domains (`nytimes.com`, `who.com`, `nih.gov`, and `cdc.com`). In contrast, the right column shows seeds scored by a different QP combination $rl, \overline{lk}, \overline{ge}_a, rt$ with 0 overlap from less popular and international (due to \overline{ge}_a) domains (e.g, `bylinetimes.com`, `rappler.com`, `thejakartapost.com`, `fiverr.com`). Table 38 also illustrates the point that different combinations of QPs result in different overlap and precision values. This point is further amplified by Figure 44 which presents the overlap (X-axis) and P@20 for 2,049 1-, 2-, and 3-combinations of Quality Proxies. The table shows how different combinations can result in high (e.g., dp, ge_a, re_b and $ge_d, re_b, \overline{re}_n$) or low ($\overline{rl}, ge_d, \overline{re}_n$) overlap/P@20. Unsurprisingly, the QP combination $\overline{rl}, ge_d, \overline{re}_n$ resulted in a low P@20 because the *relevance* rl QP was flipped, meaning relevance was penalized.

Across all topics, for Micro-collections M_G seeds, with Google seeds as the reference, the MMM average precision⁴ were 0.13, 0.58, and 0.63, respectively, when QP scores were used. Without the utilization of QP scores, the MMM were smaller; 0.06, 0.36, and 0.57, respectively. These results showed that the utilization of Quality Proxies to score seeds improved the precision of seeds by over a factor of 1.5 (0.58 vs 0.36). Also, Google seeds had the highest MMM (0.72, 0.79, and 0.94) average precision values.

²Average was calculated across the top 10 (out of 2,049) overlap scores of the QP combinations

³There were multiple QP combinations that resulted in overlap of 1.0.

⁴Average was calculated across the top 10 (out of 2,049) overlap scores of the QP combinations

Overlap and P@10 for Expert and Micro-collection seeds

Across all topics, for Expert and Micro-collection seeds, MMM average overlap were 0.25, 1.0, and 1.0, respectively, when Quality Proxies were used to score seeds. Without the utilization of QP scores, they were smaller, 0.13, 0.15, and 0.19, respectively. Similar to the overlap between Google and Micro-collection seeds (GM), these results (Table 37 and Appendix A) suggest that the utilization of QP scores facilitated the selection of seeds from a common set of domains for Micro-collection and Expert seeds.

Across all topics, for Micro-collection M_E seeds, with Expert seeds as the reference, the MMM average precision were 0.0, 0.39, and 0.95, respectively. Further investigation of the seeds that generated 0.0 precision showed that 5/10 were actually relevant based on human judgment. This means our relevance threshold of 0.20 was set too high, and thus resulted in the production of false positive labels. We discuss this problem further in Chapter 9.4.2. The MMM of the average precision of seeds not scored with QPs (M_{Er}^r) were smaller (0.06/0.20/0.61) compared to M_E by a factor of 1.95 (0.39 vs 0.20) suggesting again (as previously seen for M_G) that the utilization of QP scores improved the precision of seeds. Unlike Google seeds, the gap between the precision of Micro-collection seeds and Expert seeds was smaller; the MMM of the average precision of E was 0.13, 0.69, and 0.80, respectively.

We showed the improvement achieved as a result of using the Quality Proxies to score seeds by showing the average overlap and P@10 of Micro-collection seeds in comparison to collections that did not use Quality Proxies. However, what happens if we consider more than 10 seeds? We expect the overlap to drop since results were sorted in descending order of overlap values (Chapter 9.2.1, step 5), but how would the precision fair when more than 10 seeds, e.g., 20 or 30 seeds are considered? Table 39 (for *2020 Coronavirus Pandemic*) and Appendix B attempts to address these questions. Table 39 is similar to Table 37; the last row of averages of Table 39 is the first row of Table 37 since $K = 10$. Table 39 shows the averages of the top 10 overlap between reference seeds and Micro-collections, and the respective P@K of the seeds for different values of K (e.g., $K = 10, 20, 100$). The last row of the table shows the result of selecting all seeds ($K = \text{All}$) irrespective of their QP scores.

Overlap and P@K for reference and Micro-collection seeds

Unsurprisingly, reference seeds from Google and Expert collections produced seeds of a higher precision than Micro-collections, reflecting the reputation of the Google SERP for

TABLE 39: (Chapter 9.3.1, Coronavirus, Supplements Table 37 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Figure 45 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - Expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of Expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

K	Average Overlap				Average P@K									
	GM	EM	G^r	M^r	E^r	M^r	G	E	G^r	E^r	M_G	M_E	M_G^r	M_E^r
10	1.0	1.0	.20	.19	.80	.80	.68	.73	.61	.95	.57	.61		
20	.80	-	.16	.15	.58	-	.67	.70	.63	-	.56	.57		
30	.56	1.0	.15	.13	.70	.76	.72	.63	.67	.79	.55	.59		
40	.44	.79	.15	.15	.70	.76	.69	.70	.77	.74	.57	.62		
50	.44	.81	.14	.13	.72	.74	.70	.67	.78	.70	.55	.58		
60	.46	.66	.15	.13	.70	.72	.69	.71	.75	.69	.57	.58		
70	.41	.64	.15	.13	.69	.76	.68	.69	.72	.71	.56	.55		
80	.37	.60	.15	.12	.69	.77	.69	.66	.71	.72	.55	.55		
90	.31	.60	.15	.12	.71	.78	.67	.70	.72	.69	.54	.55		
100	.27	.59	.15	.13	.71	.79	.70	.67	.69	.68	.57	.56		
150	.20	.58	.15	.12	.68	.82	.69	.66	.69	.72	.53	.55		
200	-	.46	-	.11	-	.81	-	.67	-	.66	-	.56		
300	-	.29	-	.10	-	.77	-	.67	-	.60	-	.55		
All	.11	.07	.11	.07	.68	.65	.68	.65	.55	.55	.55	.55		

producing quality documents - and the fact that Expert seeds were hand-selected. However, as Table 40 shows, collections of seeds from Micro-collections were all above the relevance threshold (Table 35) except *2018 World Cup* (Twitter-Latest, Table 40, No. 4). The last row of Table 39 (for *2020 Coronavirus Pandemic*) shows the average overlap and average precision of all reference (Google - G, Expert - E) and Micro-collections (M_G and M_E) seeds.

TABLE 40: Precision for reference (Google - G, Expert - E) and Micro-collections (M) seeds. Reference collections G, E produced seeds of a higher precision than M. M seeds were all above the relevance threshold (Table 35) except M from *2018 World Cup* - Twitter-Latest (*). These values were populated from the last rows of all the tables in Appendix B and Table 39.

#	Topic	Vertical	G	E	M
1	2020 Coronavirus Pandemic	Top	0.68	0.65	0.55
2	2020 Coronavirus Pandemic	Latest	0.68	0.65	0.55
3	2018 World Cup	Top	0.62	NA	0.44
4	2018 World Cup	Latest	0.62	NA	0.15*
5	Hurricane Harvey (collected 2020)	Top	0.41	NA	0.10
6	Hurricane Harvey (collected 2017)	Top	0.72	0.25	0.15
7	Flint Water Crisis	Top	0.82	NA	0.48
8	2014 Ebola Virus Outbreak	Top	0.78	0.60	0.24

This represents the baseline precision of all seeds. For example, the baseline precision of G seeds was 0.68, for E seeds it was 0.65, and for M_G and M_E it was 0.55. The baseline precision of Micro-collection seeds (0.55) was calculated without the selection of seeds with their QP scores. Previously, we saw that selecting the top 10 seeds with the highest QP scores improved the precision of seeds by a factor of 1.5 and 1.95 for M_G and M_E respectively. So we investigated if this improvement was maintained even as overlap dropped and K (number of seeds selected) was increased. Specifically, we checked if the precision of M_G and M_E could be improved beyond the baseline of 0.55 (for *2020 Coronavirus Pandemic*) when we increased K. From Table 39 and its corresponding line chart, Figure 45 (first column), as K increased, the average P@K for M_G and M_E mostly held steadily, with median of 0.71 (Standard Deviation, $\sigma = 0.05$) and 0.70 ($\sigma = 0.08$), respectively, when the QP scores were used. This was a 0.16 (M_G) and 0.15 (M_E) increase above the baseline precision (0.55). For example, the P@10 for M_G and M_E from the Table 39 were 0.61 and 0.95, respectively, and their P@100 were 0.69 and 0.68, respectively. Without the utilization of QP scores the improvement was marginal (0.01). For M_G and M_E seeds from the Twitter-Latest vertical, the P@K for M_G and M_E also maintained steadily at a median of 0.71 ($\sigma = 0.03$) and 0.75 ($\sigma = 0.05$), respectively, when the QP scores were used. This was a 0.16 (M_G) and 0.20

(M_E) increase above the baseline precision (0.55) unlike the marginal improvement (≤ 0.01) derived from the non-utilization of QP scores.

In addition to the *2020 Coronavirus Pandemic*, for other evaluation dataset topics, as K increased and overlap dropped, the P@K for M_G and M_E mostly held steady.

- For *2018 World Cup* (Table 52, Figure 46, first column), the median P@K for M_G , $MP(M_G)$ was 0.53 ($\sigma = 0.11$), a 0.09 increase above baseline precision (0.44): 0.09⁺, without QP scores, it was 0.45, a marginal improvement ($\sigma = 0.02$, 0.01⁺) above the baseline (0.44).
- For *Hurricane Harvey* (collected 2017) (Table 55, Figure 47, second column), $MP(M_G) = 0.34$ ($\sigma = 0.14$, 0.19⁺), without QP scores, it was 0.18, a marginal improvement ($\sigma = 0.03$, 0.03⁺) above the baseline (0.15).
- For *Flint Water Crisis* (Table 56, Figure 48, first column), $MP(M_G) = 0.60$ ($\sigma = 0.15$, 0.15⁺), without QP scores, $MP(M_G) = 0.50$ ($\sigma = 0.02$, 0.05⁺), a marginal improvement above the baseline (0.45).
- For *2014 Ebola Virus Outbreak* (Table 57, Figure 48, second column), with and without the use of QP scores led to a marginal improvement (≤ 0.01) above the baseline precision (0.24).

There were three cases (out of 11 total⁵) in which the utilization of QP scores reduced the median P@K as K increased, namely *Hurricane Harvey* and *2018 World Cup* Twitter-Latest seeds. This might be attributed to the fact that these seeds had the lowest Precision (Table 40), and thus, the QP scores could not improve already poor-performing seeds:

- For *Hurricane Harvey* (collected 2020) (Table 54, Figure 47, first column), $MP(M_G) = 0.09$ ($\sigma = 0.01$, -0.01⁺), without QP scores, there was no improvement above the baseline precision, similarly. For Hurricane Harvey (collected 2017), $MP(M_E) = 0.12$ ($\sigma = 0.12$, -0.03⁺), without QP scores, there was no improvement above the baseline precision (0.10).
- For *2018 World Cup* (Twitter-Latest) (Table 53, Figure 46, second column), $MP(M_G) = 0.15$ ($\sigma = 0.08$, -0.02⁺), without QP scores, there was no improvement above the baseline precision (0.15), similarly.

⁵The total includes counts of M_G and M_E seeds for each topic

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	.59	.51	.77	.74	.74	.81	.73	.88	.68	.66	.60	.52	.58	.45	.54	.56	.66	.57
20	.57	.43	.68	.63	.71	.77	.71	.72	.63	.72	.62	.63	.56	.67	.59	.73	.58	.62
30	.70	.46	.62	.56	.73	.74	.68	.66	.64	.68	.68	.71	.69	.67	-	-	-	-
40	.72	.46	.62	.54	.72	.71	.70	.66	.68	.73	.69	.77	-	-	-	-	-	-
50	.72	.45	.62	.52	.72	.72	.70	.70	.70	.64	.70	.76	-	-	-	-	-	-
60	.72	.47	.65	.51	.72	.70	.69	.72	.69	.65	.70	.76	-	-	-	-	-	-
70	.71	.48	.68	.49	.72	.71	.69	.72	.68	.65	.69	.72	-	-	-	-	-	-
80	.71	.50	.70	.48	.72	.71	.67	.70	.69	.70	.66	.72	-	-	-	-	-	-
90	.70	.51	.71	.49	.71	.70	.66	.70	.69	.72	-	-	-	-	-	-	-	-
100	.70	.49	.71	.50	.70	.69	.67	.70	.71	.70	-	-	-	-	-	-	-	-
150	.70	.54	.70	.49	.69	.68	.68	.69	-	-	-	-	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	738	724	579	578	302	288	201	193	45	43	58	58	29	29	11	11	61	61
20	413	399	748	730	567	561	116	115	55	55	41	41	42	42	28	28	39	39
30	239	227	897	872	604	602	178	178	103	103	16	16	12	12	-	-	-	-
40	205	193	908	881	686	686	187	187	49	49	14	14	-	-	-	-	-	-
50	175	163	897	870	725	725	187	187	46	46	19	19	-	-	-	-	-	-
60	133	121	880	853	783	783	187	187	48	48	18	18	-	-	-	-	-	-
70	98	88	920	894	804	804	177	177	46	46	4	4	-	-	-	-	-	-
80	67	64	980	969	815	815	166	166	18	18	3	3	-	-	-	-	-	-
90	54	54	1,043	1,043	813	813	133	133	6	6	-	-	-	-	-	-	-	-
100	41	41	1,088	1,088	793	793	126	126	1	1	-	-	-	-	-	-	-	-
150	6	6	1,073	1,073	969	969	1	1	-	-	-	-	-	-	-	-	-	-

TABLE 41: (Chapter 9.3.1 & 9.3.2, Coronavirus, Supplements Table 37 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

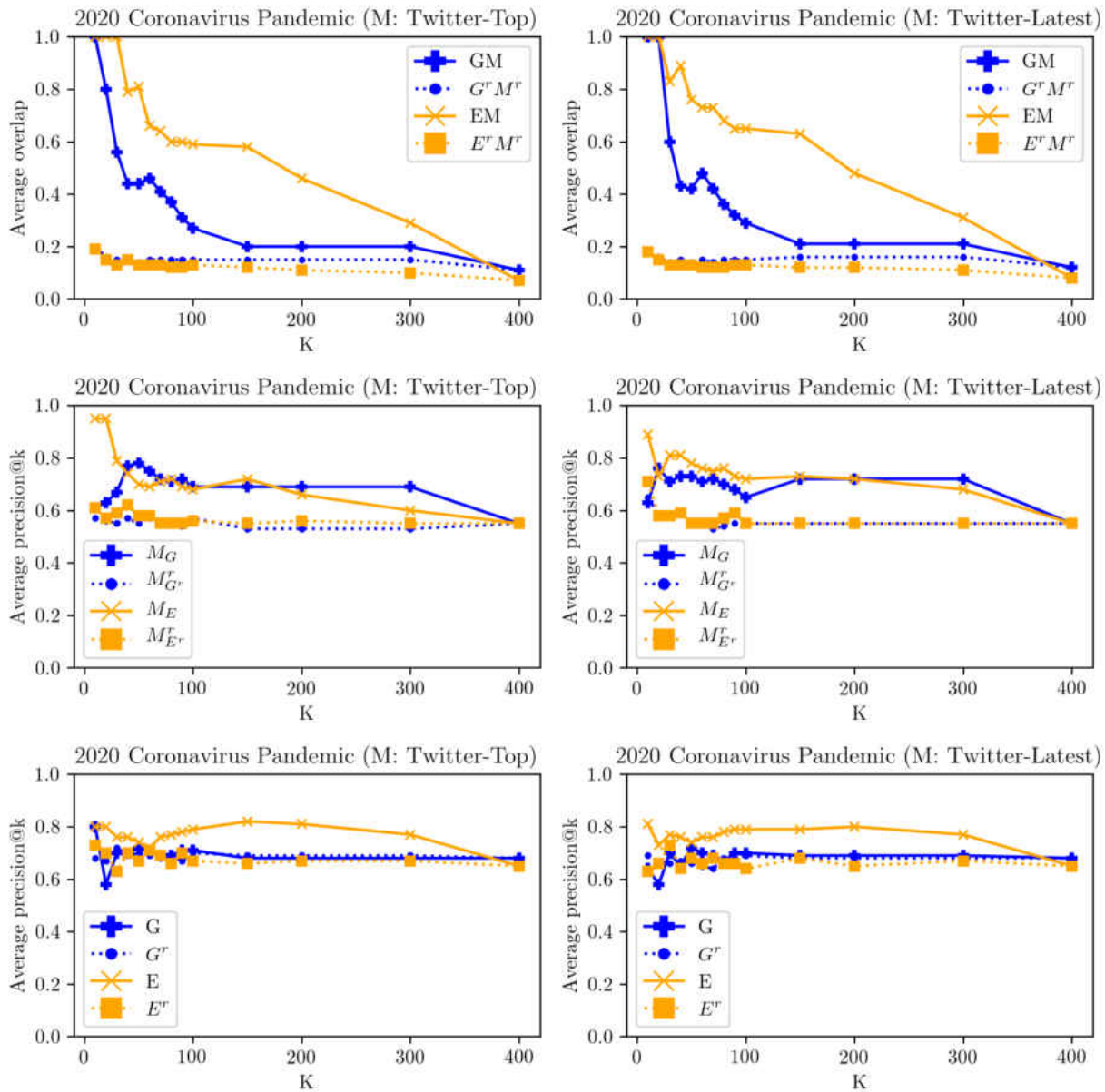


Fig. 45: (Chapter 9.3.1, Coronavirus, Supplements Table 39 & 51): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For Twitter-Top (first column), as K increased and overlap dropped, the average P@K for M_G and M_E (solid lines) mostly held steadily, with median of 0.71 ($\sigma = 0.05$) and 0.70 ($\sigma = 0.08$), respectively, a 0.16 (M_G) and 0.15 (M_E) increase above the baseline (did not use QP scores) precision (0.55). Similarly, for Twitter-Latest (second column), as K increased, the average P@K for M_G and M_E (solid lines) mostly held steadily, with median of 0.71 ($\sigma = 0.03$) and 0.75 ($\sigma = 0.05$), respectively, a 0.16 (M_G) and 0.20 (M_E) increase above the baseline precision (0.55) which did not use QP scores.

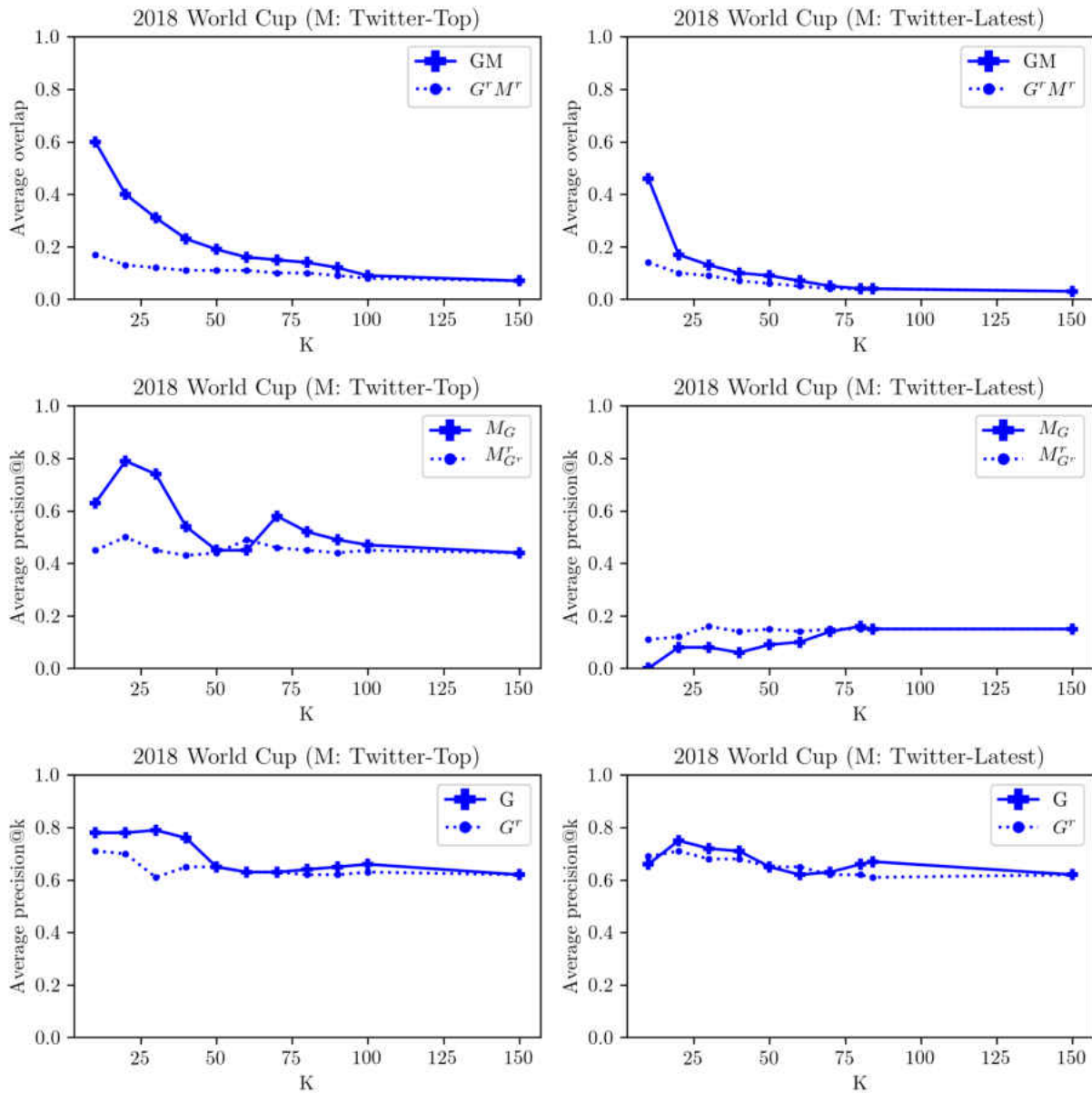


Fig. 46: (Chapter 9.3.1, 2018 World Cup, Supplements Table 52 & 53): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For Twitter-Top (first column), as K increased and overlap dropped, the average P@K for M_G (solid line) mostly held steadily, with median of 0.53 ($\sigma = 0.11$), a 0.09 increase above the baseline (did not use QP scores) precision (0.44). In contrast, for Twitter-Latest (second column), the utilization of QP scores did not improve the median P@K as K increased which might be attributed to the fact that the seeds came from a collection with the second lowest median average P@K, and thus, the QP scores could not improve already poor-performing seeds.

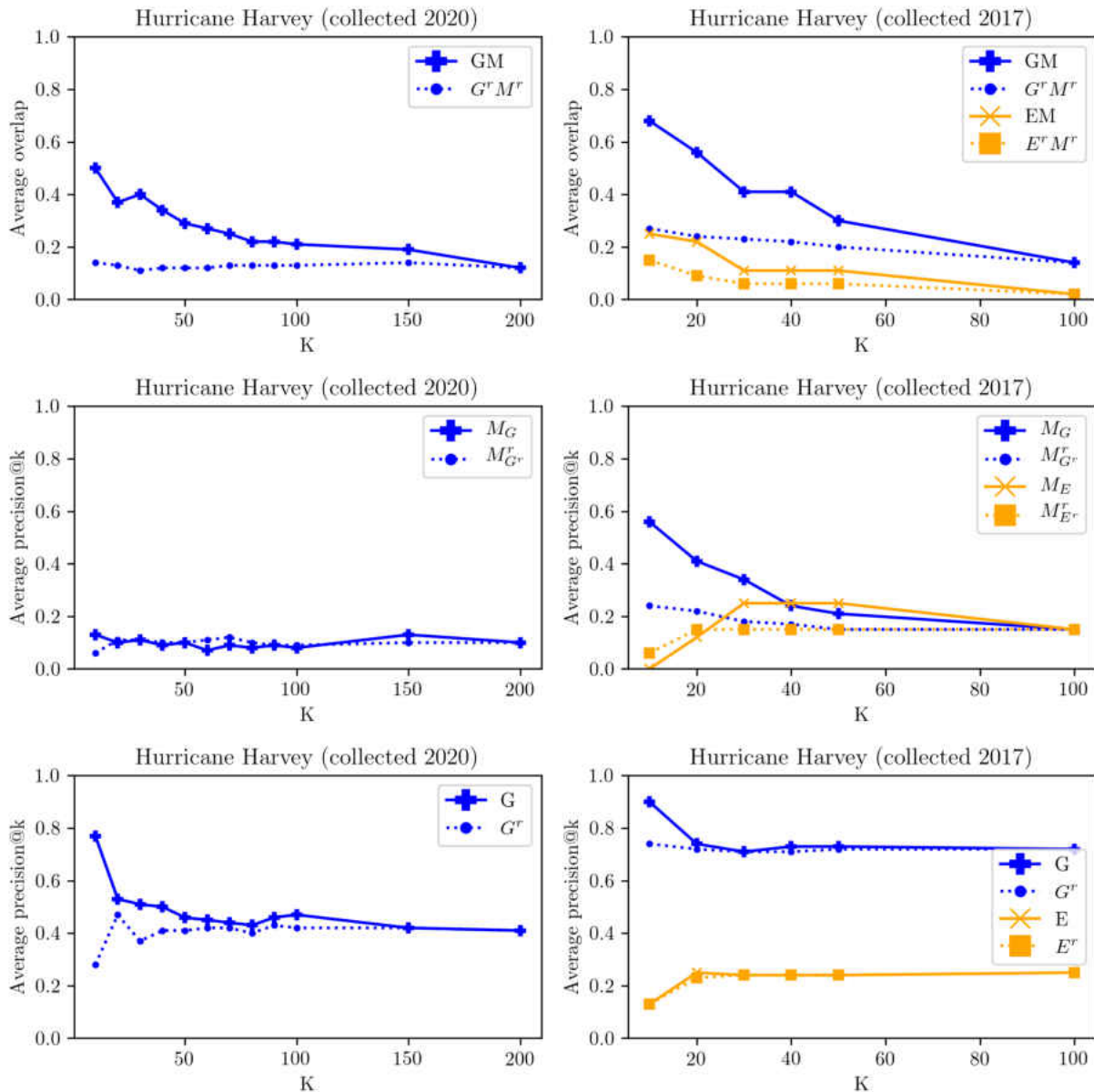


Fig. 47: (Chapter 9.3.1, Hurricane Harvey (collected 2020/2017), Supplements Table 54 & 55): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For the collection collected in 2020 (first column), three years after the event, the utilization of QP scores did not improve the median P@K as K increased, which might be attributed to the fact that the seeds came from a collection with the lowest median average P@K, and thus, the QP scores could not improve already poor-performing seeds. In contrast, for the collection collected in 2017 (second column), as K increased and overlap dropped, the average P@K for M_G (solid lines) mostly held steadily, with median of 0.34 ($\sigma = 0.14$), a 0.19 increase above the baseline (did not use QP scores) precision (0.15). However, unlike M_G , the utilization of QP scores did not improve M_E .

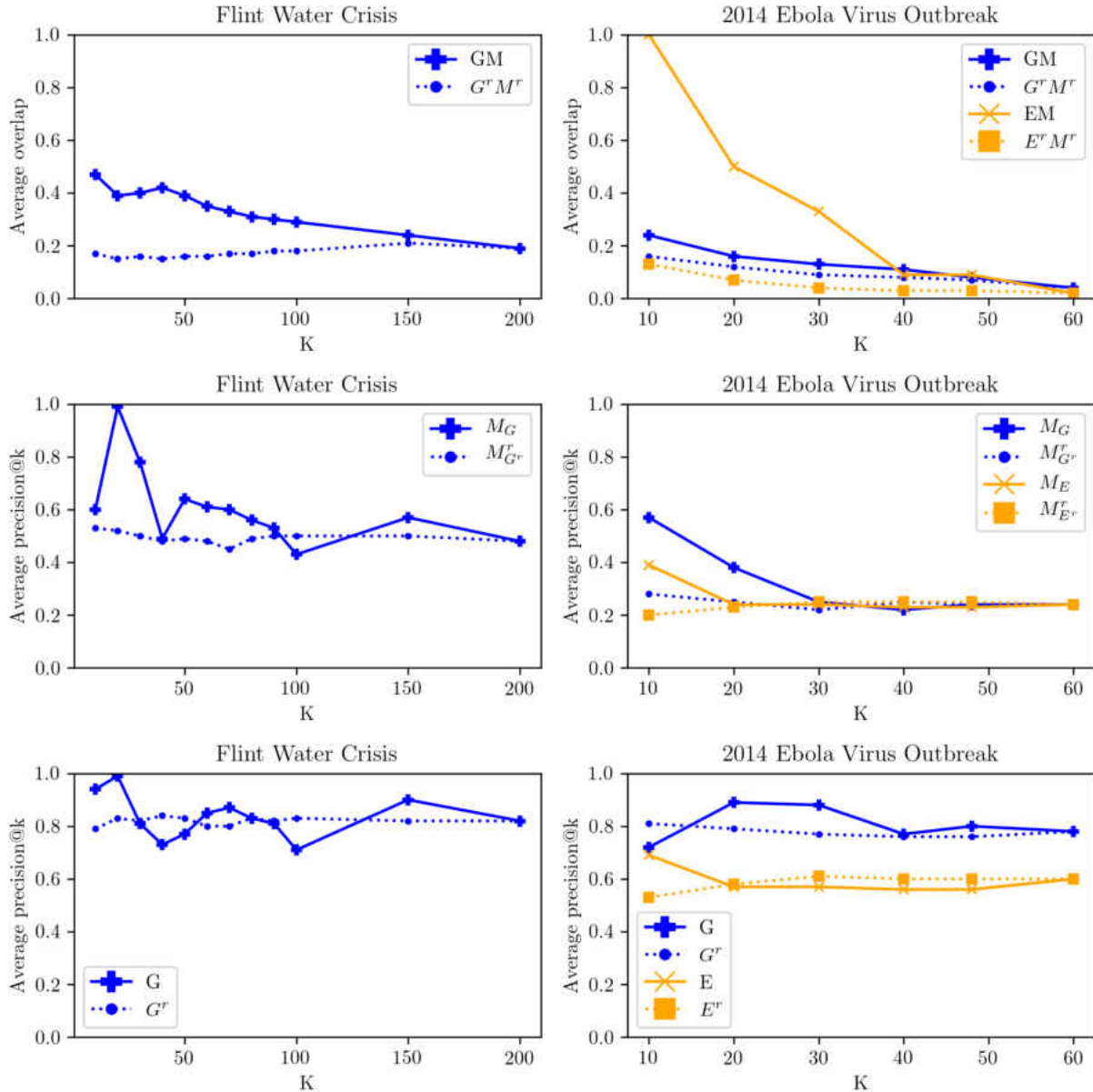


Fig. 48: (Chapter 9.3.1, Flint Water Crisis & 2014 Ebola Virus Outbreak, Supplements Table 56 & 57): Overlap (row 1) and P@K (rows 2 & 3) for K top seeds selected by their QP scores for reference Google (G)/Expert (E) seeds and Micro-collection (M) seeds. For *Flint Water Crisis* (first column), as K increased and overlap dropped, the average P@K for M_G (solid line) mostly held steadily, with median of 0.60 ($\sigma = 0.15$), a 0.15 increase above the baseline (did not use QP scores) precision (0.45). For *2014 Ebola Virus Outbreak* (second column), with and without the use of QP scores led to a marginal improvement (≤ 0.01) above the baseline precision (0.24).

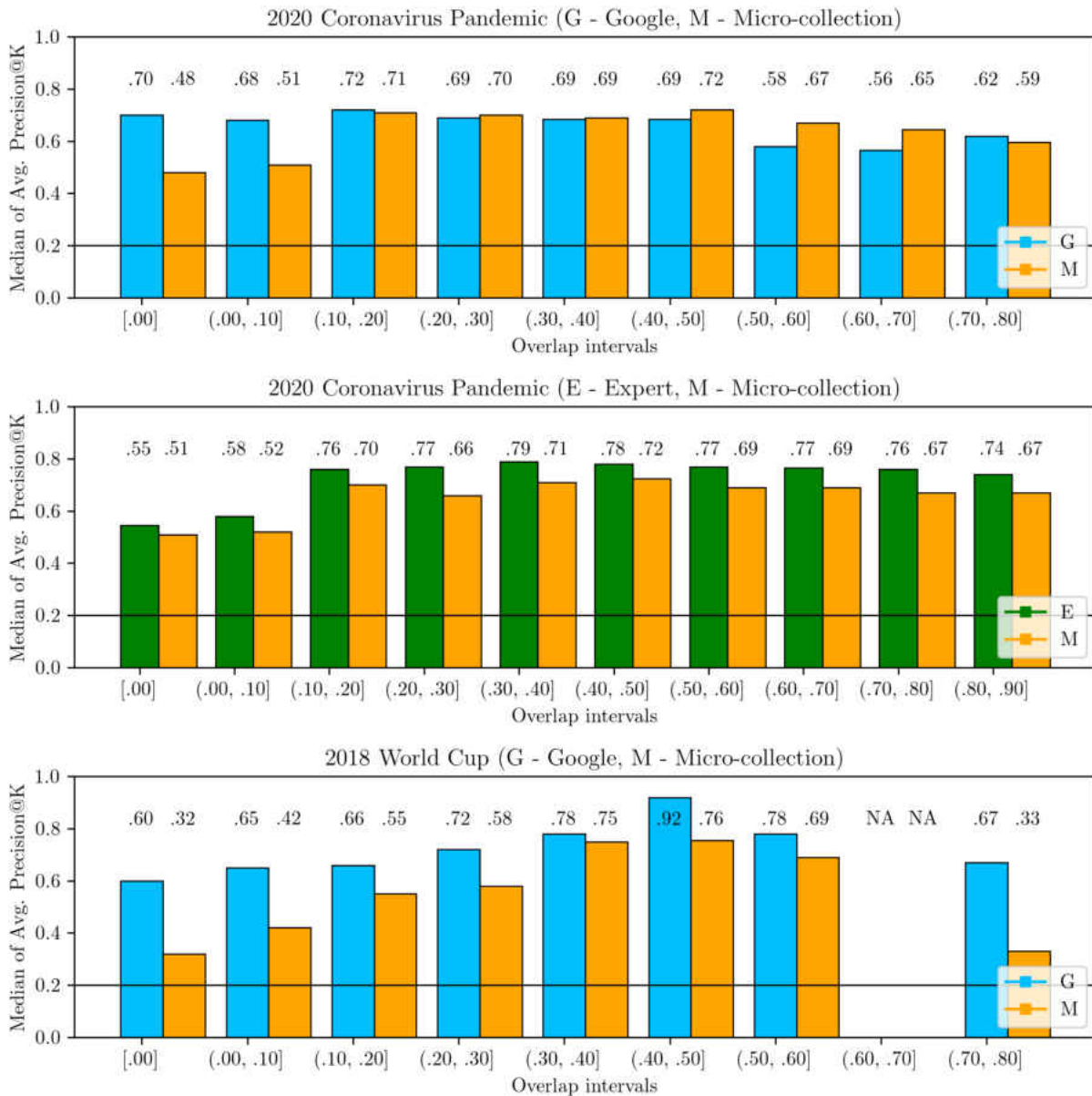


Fig. 49: (Chapter 9.3.1, Coronavirus & 2018 World Cup, Supplements Table 41 (Coronavirus) and Table 59 (2018 World Cup) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., row 1, Coronavirus - G/M, overlap of 0; P@K = 0.70/0.48) or high overlap (e.g., Coronavirus, overlap (0.70, 0.80], P@K = 0.62/0.59) between K top seeds (scored by QP scores). The black line (0.20) marks the relevance threshold.

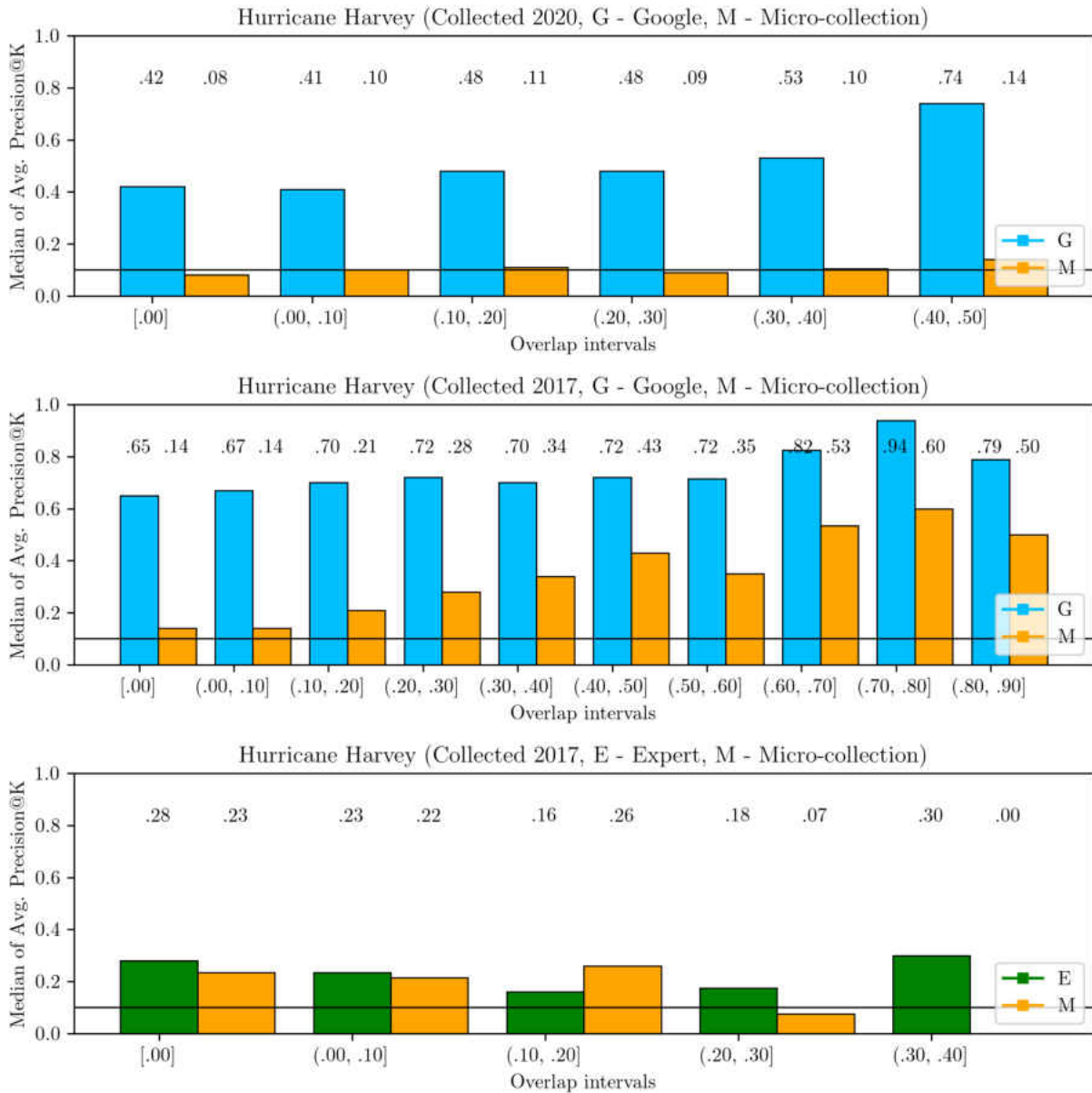


Fig. 50: (Chapter 9.3.1, Hurricane Harvey (collected 2020 and 2017), Supplements Table 60 (collected 2017) and Table 61 (collected 2017) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., first row - G/M, overlap of 0; P@K = 0.42/0.08) or high overlap (e.g., second row, overlap 0.80 – 0.90, P@K = 0.84/0.57) between K top seeds (scored by QP scores). Intuitively the higher overlap between reference seeds (high quality) and Micro-collection, the higher P@K for Micro-collections. The first and second charts aligns the most with this intuition unlike the third. The black line (0.10) marks the relevance threshold.

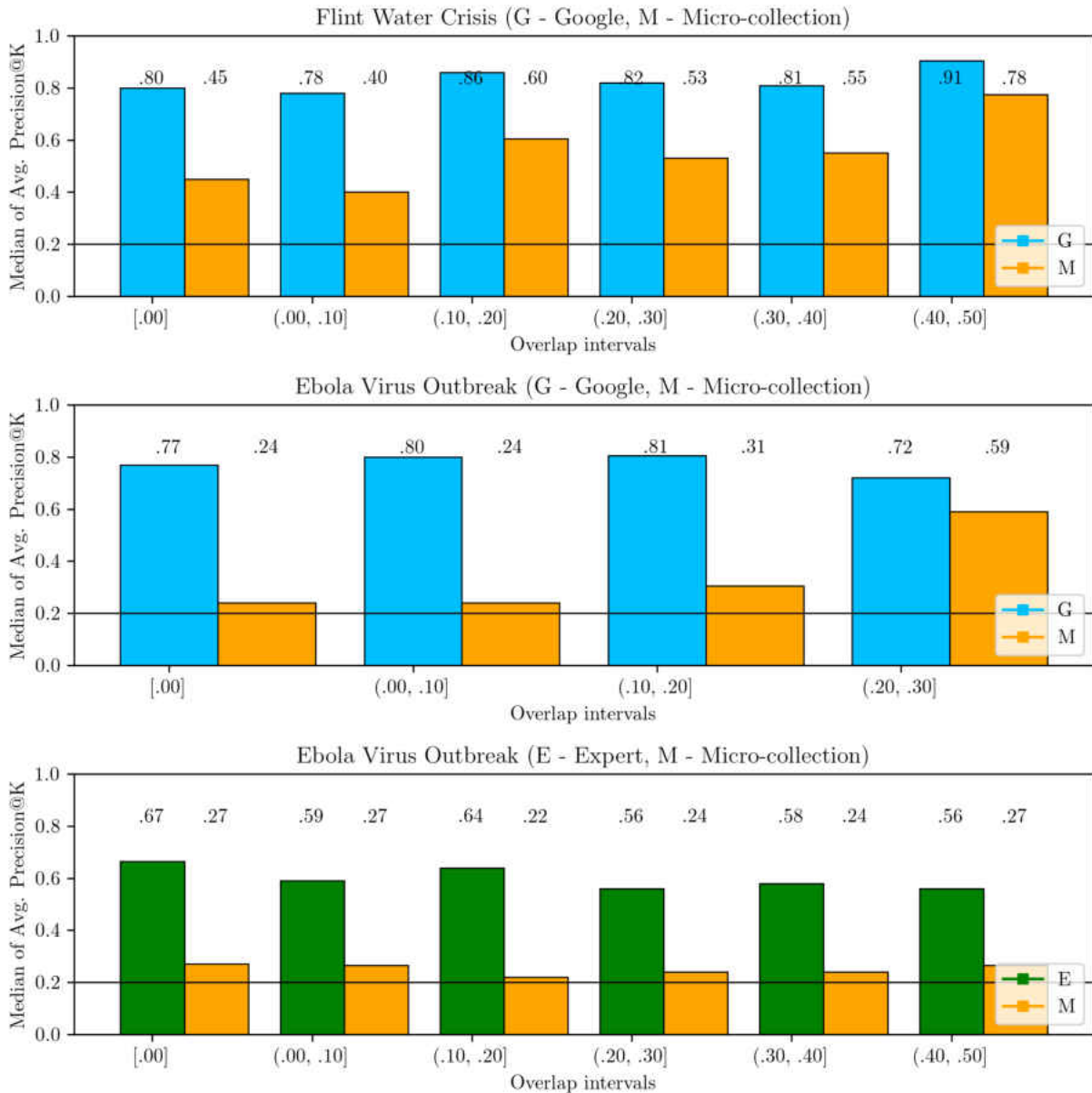


Fig. 51: (Chapter 9.3.1, Flint Water Crisis and 2014 Ebola Virus Outbreak, Supplements Table 64 (Flint Water Crisis) and Table 63 (2014 Ebola Virus Outbreak) the median of overlap intervals): P@K for overlap intervals for reference Google (G)/Expert (E) and Micro-collections (M) seeds. This distribution shows the median of average P@K for low overlap (e.g., Flint Water Crisis - G/M, overlap of 0; P@K = 0.80/0.46) or high overlap (e.g., Flint Water Crisis, overlap 0.40 - 0.50, P@K = 0.93/0.81) between K top seeds (scored by QP scores). The black line (0.20) marks the relevance threshold.

TABLE 42: 4. (Chapter 9.3.1, Coronavirus, Supplement Table 37 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combinations	Average Overlap				Average P@10							
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$
1	.22	.35	.07	.06	.77	.80	.68	.72	.89	.84	.60	.58
2	.81	1.0	.12	.13	.66	.80	.64	.65	.57	.92	.58	.65
3	1.0	1.0	.18	.17	.82	.80	.64	.64	.61	.99	.60	.54
4	1.0	1.0	.18	.19	.88	.80	.67	.73	.62	1.0	.54	.63
5	1.0	1.0	.18	.19	.91	.80	.67	.72	.66	1.0	.57	.57
6	1.0	1.0	.17	.16	.97	.80	.62	.64	.65	.98	.66	.63
7	.90	1.0	.19	.17	.90	.80	.70	.72	.57	.95	.64	.59
8	.85	1.0	.20	.18	.86	.80	.73	.64	.59	.92	.57	.62
9	.57	.67	.18	.18	.76	.80	.64	.67	.60	.93	.61	.69
10	.42	.67	.19	.20	.84	.80	.63	.64	.62	.90	.58	.57
All	.14	.20	.00	.00	.90	.80	.60	.50	1.0	1.0	.56	.50

9.3.2 RESULTS: ASSESSING SEED PRECISION WHEN NOVELTY IS PRIORITIZED

In the previous section, our results demonstrated that the utilization of Quality Proxies to score seeds resulted in the improvement of the precision of seeds selected. This was however achieved when novelty was not prioritized (overlap maximized). For example, the results presented in Table 37 (additional topics in Appendix A) and Table 39 (additional topics in Appendix B) sorted the overlap in descending order in order to examine P@K for the highest (best case) possible overlap value. Since we consider reference seeds quality seeds, we expect a high overlaps between Micro-collections and reference seeds to be associated with a high P@K for reference seeds, but this question, addressed in this section, remains - how will the P@K of Micro-collection seeds fair when they have small (high novelty) or no overlap (maximum novelty) with reference seeds? In other words, can we quantify the P@K of Micro-collection seeds when novelty (with respect to reference) is prioritized?

Table 41 - top (for *2020 Coronavirus Pandemic*) shows the average P@K for different overlap intervals unlike Table 37 which shows the top 10 ($K = 10$) overlap and P@10 for

seeds scored with QPs. The average precision was calculated for all QP combinations that produced overlap within the specified interval (e.g., interval - 0). For example, from Table 41 - top, when the top 10 seeds ($K = 10$) with the highest QP scores were selected, the average P@10 for the Google (G) and Micro-collection (M) seeds was 0.59, and 0.51, respectively. For each cell in Table 41 - top, the bottom Table shows the number of QP combinations that produced the average P@K. Table 58 - top and bottom are similar to Table 41 - top and bottom, but for Expert (E) and Micro-collection (M) seeds. Appendix C includes additional Tables of this variant for the remaining topics. Figures 49, 50, and 51 supplement Tables 41, 58, and Appendix C by visualizing the median of the average P@K for different topic.

From Figures 49, 50, and 51, in addition to the bar charts that represent the heights of the median P@K for different overlap intervals, horizontal lines mark the relevance threshold for each dataset topic. These charts reveal that in all cases except *Hurricane Harvey* (collected 2020) the median of the average P@K of M seeds for the 0 overlap (maximum novelty) interval was always above the relevance threshold. This suggests the maximum novelty (0 overlap) did not adversely affect the P@K for M seeds even though the benefit of a higher overlap varied across different topics.

2020 Coronavirus Pandemic: P@K for Micro-collection seeds when novelty is prioritized

The median (*Med*) of the average P@K (*MedK*) for Micro-collection (M) seeds with Google (G) reference seeds, for 0 overlap was 0.48, $MedK(0) = 0.48$ (Figure 49, row 1, orange-colored bars). It was 0.51 (0.03 increase) for (0, 0.10] overlap interval; $MedK(0, 0.10] = 0.51$. For the largest overlap interval, (0.70, 0.80], $MedK(0.70, 0.80] = 0.59$. This means that going from 0 overlap to (0.70, 0.80], increased the P@K by 0.11. Consequently, the P@K of Micro-collection seeds was not adversely affected when novelty was maximum (lowest overlap). Similarly, for Expert reference seeds (Figure 49, row 2), $MedK(0) = 0.51$ and $MedK(0.80, 0.90] = 0.67$. This means that going from 0 overlap to (0.80, 0.90] increased the precision by 0.16.

Hurricane Harvey: P@K for Micro-collection seeds when novelty is prioritized

Figure 50 (rows 1 and 2) contrasts two *Hurricane Harvey* seed collections. The *Hurricane Harvey* (Figure 50 row 1) collected in 2020, three years after the natural disaster, had the lowest precision (0.10, Table 40 No. 5) unlike the second collection (Figure 50 row 2) with 0.15 precision (Table 40 No. 6), which was collected the same year as the natural disaster,

TABLE 43: Median of the average overlap (\max^+ , \min^-) and P@10 for dataset topics for lower order (1 – 3) and higher order (4 – 10 and All) combinations. The combination *All* means that all Quality Proxies (without flipped state) were used to score the seeds.

Combination	Average Overlap			Average P@K					
	GM	EM	$\frac{GM+EM}{2}$	M_G	G	$\frac{M_G+G}{2}$	M_E	E	$\frac{M_E+E}{2}$
1	0.21	0.35	0.28	0.49	0.77	0.63⁻	0.47	0.73	0.60
2	0.43	1.00	0.71	0.51	0.79	0.65	0.39	0.69	0.54
3	0.54	1.00	0.77⁺	0.59	0.80	0.69	0.28	0.71	0.49⁻
4	0.43	1.00	0.71	0.58	0.87	0.72	0.50	0.70	0.60
5	0.45	1.00	0.72	0.60	0.90	0.75⁺	0.50	0.70	0.60
6	0.43	1.00	0.71	0.59	0.90	0.74	0.50	0.70	0.60
7	0.39	1.00	0.69	0.53	0.87	0.70	0.50	0.70	0.60
8	0.41	1.00	0.70	0.56	0.81	0.68	0.50	0.70	0.60
9	0.35	0.67	0.51	0.57	0.79	0.68	0.50	0.72	0.61
10	0.34	0.67	0.50	0.56	0.81	0.68	0.50	0.74	0.62
<i>All</i> (12)	0.13	0.20	0.16⁻	0.47	0.79	0.63⁻	0.50	0.80	0.65⁺

suggesting the importance of collecting seeds early for events with well-defined start and end durations. The P@K of the *Hurricane Harvey* M seeds (collected 2017, Figure 50, row 2, orange-colored bars), benefited the most from increasing overlap. Going from overlap of 0 to (0.80, 0.90] overlap increased *MedK* by 0.36 (from 0.14 to 0.50). Also, even though its $MedK(0) = 0.14$ was low compared to $MedK(0.80, 0.90] = 0.50$, 0.14 was above the relevance threshold (0.10, Table 35, No. 1).

9.3.3 RESULTS: ASSESSING THE QUALITY OF HIGHER ORDER COMBINATIONS VS. LOWER ORDER COMBINATIONS

Thus far, we have reported overlap and P@K results from scoring seeds with lower order (1, 2, and 3) QP combinations (e.g., Table 36). What if we used higher order (3+) combinations, e.g., 4, 5, 6, or all the QPs to score seeds, would their overlap and/or P@K outperform lower order combinations? Addressing this question is critical for two primary reasons. First, if it is shown that lower order combinations perform approximately the same or better than higher order combinations, this is positive news because it is computationally

cheaper to generate and utilize lower order QP combinations than higher order combinations. Second, if it is shown that higher order combinations perform better than the already well-performing lower order combinations, this means we can further improve the quality of seeds by increasing the overlap and/or P@K.

Figures 52, 53, and 54 present the overlap and P@K for different combinations across different topics. Table 43 consolidates the overlap and P@K irrespective of the topic by showing the median of the average overlap and P@K for each lower order and higher order combinations. From this table we see that the highest overlap was achieved with a lower order 3-combination QPs (GM = 0.54 and EM = 1.0). Surprisingly, the least performing combination was the utilization of all the QPs (GM = 0.13 and EM = 0.20) to score the seeds. This indicates more QPs does not necessarily increase overlap.

For P@K, the result was a mixed bag, although overall higher-order higher order combinations produced higher P@K values. 5-combinations produced the highest P@K for M_G and G seeds, while the lowest-order (1-combination) and a higher-order *All* QP combination produced the worst P@K values. However, for M_E and E, the best performing combination was the utilization of *All* QPs (M_E : 0.50 and E: 0.80), meaning these seeds benefited from higher order combinations unlike M_G and G.

9.3.4 RESULTS: ASSESSING DIVERSITY OF SEEDS FROM MICRO-COLLECTIONS

All figures in Appendix F present the CDFs of the diversity of reference (Google - G, Expert - E) and Micro-collection seeds as well as the following variants: diversity of Micro-collection seeds with Google seeds as reference, M_G , and Expert as reference, M_E . Other variants include the diversity of seeds selected without the utilization of Quality Proxy scores. These seeds (G^r , E^r , M_G^r , M_E^r , etc) have the r -superscript. Table 44 summarizes the CDF figures in Appendix F by reporting the diversity d_u of seeds for their Quartiles (Q1, Q2, and Q3). Similarly, Table 45 summarizes the CDF figures in Appendix G by reporting the diversity d_c (Chapter 9.2.3) of the seeds for their quartiles. Both tables are read similarly, for example, from Table 44, Collection No. 1. (*2020 Coronavirus (Top)*), 25% of the Google (G) seeds had diversity ≤ 0.60 , the median (Q2) was 0.80, and 75% (Q3) of the Google seeds had diversity ≤ 1.0 . The findings from using both diversity measures d_u (Table 44) and d_c (Table 45) were consistent even though the magnitudes of the specific diversity values differ. Consequently, we only reported d_u to summarize our findings, hence, all diversity references refer to d_u .

	Q1	Q2	Q3		Q1	Q2	Q3		Q1	Q2	Q3							
1. 2020 Coronavirus (Top)								5. Hurricane Harvey (collected 2020)										
G	.60	.80	1.0	G^r	.90	1.0	1.0	G	1.0	1.0	1.0	G^r	1.0	1.0	1.0			
M_G	.50	.80	.90	M_G^r	.90	.90	1.0	M_G	.50	.80	1.0	M_G^r	.80	.90	1.0			
E	.40	.80	.90	E^r	1.0	1.0	1.0	6. Hurricane Harvey (collected 2017)										
M_E	.50	.80	.90	M_E^r	.90	.90	1.0	G	.60	.70	.90	G^r	.80	.80	.90			
2. 2020 Coronavirus (Latest)								M_G				.70	.90	.90	M_G^r	.90	1.0	1.0
G	.60	.80	1.0	G^r	1.0	1.0	1.0	E	.40	.50	.70	E^r	.60	.70	.80			
M_G	.50	.80	1.0	M_G^r	.90	.90	1.0	M_E	.70	.90	.90	M_E^r	.90	1.0	1.0			
E	.40	.80	1.0	E^r	1.0	1.0	1.0	7. Flint Water Crisis										
M_E	.60	.80	1.0	M_E^r	.80	.90	1.0	G	.90	1.0	1.0	G^r	1.0	1.0	1.0			
3. 2018 World Cup (Top)								M_G				.50	.80	.90	M_G^r	.90	1.0	1.0
G	.80	1.0	1.0	G^r	.90	1.0	1.0	8. 2014 Ebola Virus Outbreak										
M_G	.40	.60	.90	M_G^r	.80	.90	.90	G	.90	1.0	1.0	G^r	1.0	1.0	1.0			
4. 2018 World Cup (Latest)								M_E				.20	.50	.70	M^r	.50	.60	.70
G	.80	1.0	1.0	G^r	.90	1.0	1.0	E	.50	.60	.80	E^r	.80	.90	.90			
M_G	.50	.70	.90	M_G^r	.70	.80	.90	M_E	.20	.50	.70	M^r	.50	.60	.70			

TABLE 44: First (Q1), second (Q2), and third (Q3) quartiles of Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of reference (Google - **G** & Expert - **E**) and Micro-collection (M_E & M_G) seeds. A single cell, e.g., **G**, **Q1**, from **No. 1**, reads as follows 25% of seeds had diversity \leq **0.60**. Overall, seeds (with r -superscript) selected without using QP scores has a higher diversity of seeds selected with QP scores, Google seeds had the highest diversity, while the diversity of Micro-collection and Experts seeds was similar. **Key:** **green** - column-wise maximum, **red** - column-wise minimum.

	Q1	Q2	Q3		Q1	Q2	Q3		Q1	Q2	Q3							
1. 2020 Coronavirus (Top)								5. Hurricane Harvey (collected 2020)										
G	.62	.66	.70	G^r	.70	.71	.72	G	.67	.69	.70	G^r	.70	.71	.73			
M_G	.57	.65	.69	M_G^r	.67	.69	.71	M_G	.58	.65	.70	M_G^r	.66	.69	.70			
E	.55	.64	.70	E^r	.70	.72	.72	6. Hurricane Harvey (collected 2017)										
M_E	.57	.65	.69	M_E^r	.67	.69	.71	G	.58	.63	.68	G^r	.64	.66	.68			
2. 2020 Coronavirus (Latest)								7. Flint Water Crisis										
G	.62	.66	.70	G^r	.70	.71	.73	G	.66	.69	.72	G^r	.70	.71	.73			
M_G	.60	.66	.71	M_G^r	.67	.70	.71	M_G	.63	.66	.71	M_G^r	.69	.70	.72			
E	.55	.64	.71	E^r	.70	.72	.73	E	.52	.59	.64	E^r	.61	.64	.66			
M_E	.60	.66	.71	M_E^r	.67	.69	.71	M_E	.62	.66	.71	M_E^r	.68	.70	.72			
3. 2018 World Cup (Top)								8. 2014 Ebola Virus Outbreak										
G	.66	.69	.72	G^r	.69	.71	.72	G	.68	.71	.73	G^r	.70	.72	.73			
M_G	.54	.62	.71	M_G^r	.66	.69	.71	M_G	.43	.60	.66	M_G^r	.62	.65	.68			
4. 2018 World Cup (Latest)								E				.54	.64	.69	E^r	.69	.71	.72
G	.67	.70	.72	G^r	.69	.71	.72	M_E	.43	.60	.66	M_E^r	.62	.65	.68			
M_G	.60	.69	.72	M_G^r	.66	.69	.71											

TABLE 45: First (Q1), second (Q2), and third (Q3) quartiles of Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of reference (Google - **G** & Expert - **E**) and Micro-collection (M_E & M_G) seeds. A single cell, e.g., **G**, **Q1**, from **No. 1**, reads as follows 25% of seeds had diversity \leq **0.60**. Overall, seeds (with r -superscript) selected without using QP scores has a higher diversity of seeds selected with QP scores, Google seeds had the highest diversity, while the diversity of Micro-collection and Experts seeds was similar. **Key:** **green** - column-wise maximum, **red** - column-wise minimum.

Overall, the median diversity of Micro-collection seeds (M_E and M_G) was between 0.50 and 0.90. However, seeds selected without using QPs had higher diversity than those selected with QPs. This was not unexpected because seeds from the same domain have similar QP scores since they share a common set of QP dimensions (dp , ge_d , sc , reb , and ren). For example, the five `nytimes.com` seeds in Table 38 have QP scores: 0.61, 0.58, 0.57, 0.54, and 0.53. Consequently, when seeds are sorted in order of their respective QP scores, seeds from a common set of domains are ranked near one another, causing a reduction in diversity (due to repetition of domains) if they are selected. However, the focus of our evaluation was to assess the diversity of reference and Micro-collection seeds and not the diversity with or without the utilization of Quality Proxy scores.

The median diversity of Micro-collection M_G seeds was lower compared to Google for all the collections except the *Hurricane Harvey (collected 2017)* collection, and the *2020 Coronavirus Pandemic* (equal median). This could be due to the fact that all the Google collections were extracted (Table 33) from Pages 1 – 10 or 20, and thus minimize the resampling of seeds from the same domain. However, when seeds from Google are collected from Page 1, this increases the probability of resampling the top ranked seeds from the same domain. This was the case for the *Hurricane Harvey (collected 2017)* collection. The median (Q2) diversity of Google G seeds was 0.80 (vs. 0.80 M_G) for the *2020 Coronavirus Pandemic*, 0.70 (vs. 0.90 for M_G) for *Hurricane Harvey (collected 2017)*, 1.0 (vs. 0.60 and 0.70 for M_G) for *2018 World Cup*, 1.0 (vs. 0.80 for M_G) for *Hurricane Harvey (collected 2020)*, 1.0 (vs. 0.80 for M_G) *Flint Water Crisis*, and 1.0 (vs. 0.50 for M_G) *2014 Ebola Virus Outbreak* collections.

Unlike M_G , the median diversity of Micro-collections M_E seeds was higher compared to seeds from Experts for one of two topics, and equal for two topics. This could be due to the fact that Micro-collections are sampled seeds posted by multiple individuals, which results to the sampling of a larger pool of domains unlike expert seeds which are created by fewer individuals. The median diversity of Micro-collection M_E seeds was 0.80 (vs. 0.80 for E) for the *2020 Coronavirus Pandemic* (Twitter-Top/Latest), 0.90 (vs. 0.50 for E) for *Hurricane Harvey (collected 2017)*, and 0.50 (vs 0.60 for E) for the *2014 Ebola Virus Outbreak* collections.

9.4 LIMITATIONS OF FRAMEWORK

The following is a description of some of the limitations we have identified for our framework.

9.4.1 ESOTERIC STORIES

The evaluation dataset topics such as the *2020 Coronavirus Pandemic* and *Flint water crisis* are well-documented stories. Over the years or the past few months⁶, social media users have posted about these stories and multiple Wikipedia editors have collaborated to document these topics. We expect that there are important stories with little or no news media coverage, for example esoteric stories, for which there are few news articles, and as a result, few or no social media posts with seeds. Such poorly covered stories would result in a shortage of seeds. Similarly, not every important news story has a canonical Wikipedia page. The absence of a Wikipedia page would affect the calculation of *reputation* (re_b and re_n) QPs. Our framework is expected to perform poorly for such esoteric stories.

9.4.2 LIMITATIONS OF QUALITY PROXIES

Having multiple QPs is advantageous since it provides the ability to still calculate the quality scores for seeds even when a subset of the QPs are absent. However, the Quality Proxies have some limitations as described below.

Limitations of Domain Popularity (dp) QP

The domain popularity dp QP relies on identifying the social media account (e.g., @CNN) associated with a domain (e.g., `cnn.com`). This information can be absent or ambiguous (e.g., @CNNBRK and @cnni all belong to `cnn.com`), posing a problem to the instantiation of the dp QP.

Limitations of Geographical (ge_a and ge_d) QPs

Geographical information can be absent, non-machine readable, ambiguous, or false. All of these pose obstacles to the instantiation of the geographical QP with high-quality information.

Limitations of Retrievability (rt) QP

The retrievability rt QP attempts to estimate how easy it is to find a seed. There are multiple ways of measuring retrievability, each with its limitations. We measured retrievability of a seeds by checking if the seeds were found within the first 10 or 20 pages of the Google SERP

⁶Time of writing reference point: October 2018 – July 2020

for a given query. This method of measuring retrievability is highly sensitive to time. The SERP of an ongoing news story constantly receiving updates is expected to be in flux as we showed in our previous work [24]. This means retrievability changes with time, and rt does not capture the dynamics of retrievability.

Limitations of Relevance (rl) QP and Precision thresholds evaluation

Measuring relevance required calculating the cosine similarity between a seed’s document vector and the gold-standard document vector. Seeds can contain small text (e.g, a webpage for loading videos), hard-to-extract text (e.g, PDFs), or no text (e.g., images). All pose problems for generating document vectors and result in a small or no similarity between a seed and a gold-standard document vector for some on-topic seeds. This often results false negative errors: the labeling of relevant seeds as non-relevant.

Estimating precision required the setting of relevance thresholds (Table 35), the similarity threshold for which seeds must reach to be considered relevant. There are multiple ways of setting this threshold. To estimate the relevance threshold, for each dataset topic, we first measured the one-vs-rest similarity between each document and the rest of the documents in the gold-standard. Second, we calculated the median of the one-vs-rest similarity and set the threshold as the median of the one-vs-rest similarity clipped at 0.20. We determined that this method resulted in setting the relevance threshold too high, resulting in a strict assessment of precision, which does not give enough room for novelty in documents. As a result of this, there were multiple false negative precision errors: the labeling of relevant documents as non-relevant.

9.5 GENERATING SEEDS WITH THE MCQP FRAMEWORK, A RECOMMENDATION

As we outlined in Chapter 9.1, generating seeds begins with issuing a query to the social media search engines, and subsequently extracting links from Micro-collections ($\mathbf{P}_n\mathbf{A}_1 \cup \mathbf{P}_n\mathbf{A}_n$). Our framework proposes extracting seeds from Micro-collections, however, the user of the framework could additionally extract seeds from non-Micro-collections ($\mathbf{P}_1\mathbf{A}_1$) since the Quality Proxies work for all post classes. Similarly, our framework excludes extracting seeds from Google, however, the user could extract seeds ($\mathbf{P}_1\mathbf{A}_1$) from Google. In Chapter 5.5, we provided a recommendation for generating seeds from social media, the first component of the MCQP framework. In this section, we propose multiple recommendations for generating seeds. First, we propose recommendations for the sources (Google and/or social

media) to consider when extracting seeds, based on the attributes prioritized (quality, quantity, hostname diversity, and age) by the curator. Second, we propose recommendations for utilizing the Quality Proxies to score and select seeds.

9.5.1 RECOMMENDATIONS FOR SEED SOURCES

The attributes prioritized by a curator inform their choice in the selection of a source to generate seeds. Attributes such as prioritizing *quality*, *quantity*, *hostname diversity*, and *age* could affect whether a curator samples seeds from Google or Twitter-Latest. Consequently, informed by this research effort, we propose the following recommendations for the sources of seeds curators should consider based on the attributes important to them.

Quality of Seeds Prioritized

A curator that prioritizes quality should consider extracting seeds first from Google-All (default vertical), second, Twitter-Top, and third, Reddit-Relevance. Based on our experience, the Google search engine produces better quality documents than social media search engines, especially the Reddit SERP, which we empirically determined to rely heavily on string-matching for query understanding.

Quantity of Seeds Prioritized

A curator that prioritizes extracting many seeds should consider extracting seeds from the following ordered-list of sources: Reddit-Comments, Reddit-Top, Twitter-Latest, Google, and Wikipedia. The curator should note that multiple factors could affect the number of unique seeds extracted from a source, such as the topic, temporal gap between sampling seeds, the age of the news story, the popularity of the news story, etc.

The Reddit-Comment SERP vertical orders post by the number of replies or comments they have. Since posts with larger comments often involve a larger number of users, comments provides the opportunity for multiple users to post links. Similarly, Reddit-Top ranks posts by there popularity, and since popular posts engage more users, popular posts provide the opportunity for multiple users to post links. However, the curator should note that these seeds might not be relevant since the Reddit-Comment and Reddit-Top verticals do not prioritize relevance unlike Reddit-Relevance.

The Twitter-Top SERP filters tweets based on some notion of popularity (e.g., combination of top retweets, likes, freshness) resulting in the same set of tweets maintaining the

same position for a given time period. This means it is possible to sample the Twitter-Top SERP multiple times (e.g., days apart) and extract the same seeds. This can also happen when collecting seeds from the Twitter-Latest vertical, however, since the Twitter-Latest vertical is in recent-first order, it receives more supply of new tweets than Twitter-Top. The curator should note that persistent (e.g., minutes apart) scraping of tweets could result in Twitter throttling the user agent, restricting or stopping the extraction of tweets. This might be avoided if the curator alternates the user agent.

To maximize extracting many seeds from Google, the curator should paginate. Similar to Twitter, the curator should note that the Google SERP throttles or blocks scrapers by issuing CAPTCHAs, besides the number of pages accessible for a given query is restricted.

Wikipedia articles do not exist for every news story or event but curators can extract seeds from Wikipedia references. The number of references can be affected by the popularity and/or age of the news story.

Hostname Diversity of Seeds Prioritized

A curator that prioritizes extracting seeds from a diverse set of hosts should consider extracting seeds from the following ordered-list of sources: Reddit, Twitter, and Google. In general, social media sources (e.g., Reddit and Twitter) produce seeds with a higher domain diversity than Web search engines. Domain diversity from Google can be maximized by extracting seeds from more SERPs. Across all sources, diversity can be maximized by collecting seeds persistently (e.g., daily or weekly).

Age of Seeds Prioritized

A curator that prioritizes extracting older seeds should consider extracting seeds from the following ordered-list sources that do not order posts by recency: Wikipedia, Reddit-Comments, Reddit-Top, Reddit-Relevance, and Google. In contrast, a curator that prioritizes extracting newer seeds should consider extracting seeds from the following ordered-list of sources that order posts by recency: Google, Twitter-Latest, and Reddit-New.

9.5.2 RECOMMENDATIONS FOR USING THE QUALITY PROXIES

Following the extraction of seeds, we recommend instantiating all possible dimensions of the Quality Proxy vector for all seeds. The more the Quality Proxy dimensions instantiated, the higher integrity of the Quality Proxy score. However, we acknowledge that this might

not always be possible, consequently the Quality Proxy score was designed to accommodate the utility of a subset or combination of Quality Proxies.

After instantiating the Quality Proxy vector for the seeds, the user could use all or a subset of the QPs to assign a quality score to the seeds and only select seeds scores exceeding a user-defined threshold, or the top K seeds. The utilization of a subset of QPs requires prior knowledge to determine the combination of QPs the suits the needs of the user. For example, in the previous chapter we showed how different combinations of Quality Proxies such as $\{ge_a, ge_d\}$, $\{rp, sh, lk\}$, $\{re_b\}$, and $\{re_n\}$ map to different policies for selecting seeds. These policies include prioritizing the selection from locals users and local news media (ge_a, ge_d), popular popular posts (rp, sh, lk), broadly (re_b) or narrowly (re_n) reputable sources and much more. It is up to the user to determine the policy and combine the Quality Proxies to approximate the policy.

9.6 CHAPTER SUMMARY

In this chapter we presented our MCQP framework for bootstrapping Web archive collections from Micro-collections in social media. Next, we assessed the quality of seeds selected through the utilization of Quality Proxy scores in two primary ways. We measured the P@K of seeds when novelty (approximated with overlap) was and was not prioritized. Additionally, we measured the diversity of Micro-collection seeds with respect to reference seeds selected with their respective QP scores.

Our results for $K = 10$ suggest that the utilization of QP scores helped surface seeds from a common set of domains for Micro-collection and Google. Across all topics, for Google and Micro-collection seeds, the Minimum, Median, and Maximum (MMM) average overlap were 0.24, 0.55, and 1.0, respectively, when Quality Proxies were used to score seeds. Without the utilization of QP scores, the MMM average overlap were smaller - 0.14, 0.17, and 0.27, respectively.

Similarly, for Expert and Micro-collection seeds, the utilization of QP scores facilitated the selection of seeds from a common set of domains for Micro-collection and Expert seeds. Across all topics the MMM average overlap were 0.25, 1.0, and 1.0, respectively, when Quality Proxies were used to score seeds. Without the utilization of QP scores, they were smaller - 0.13, 0.15, and 0.19, respectively.

Additionally, still for $K = 10$, our results showed that the utilization of Quality Proxies to score seeds improved the precision of Micro-collection seeds by over a factor of 1.5 (0.58 vs 0.36 median of average precision) when Google was the reference, and by a factor of

1.95 (0.39 vs 0.20 median of average precision) when Expert was the reference. For greater values of K ($20 \leq K \leq 300$, and $K = \text{all seeds}$), the median improvement in precision of Micro-collection seeds when Quality Proxies scores were used to select seeds 0.12 when Google was the reference and 0.07 when Expert was the reference.

We measured the P@K of seeds when novelty was prioritized by quantifying the precision of seeds when overlap was 0 (maximum novelty). Our results suggested that the quality of seeds selected by Quality Proxies was not compromised even when overlap was low; for M_G with 0 overlap, the median of the average P@K of five of six collections were above their respective relevance thresholds. Similarly, for M_E , with 0 overlap, the median of the average P@K of all the M_E collections were above their respective relevance thresholds.

Finally, diversity was impacted by how the dataset was created. Overall, using Quality Proxies reduced the diversity of seeds selected due to the resampling of top ranked seeds from a common set of domains. Also, Google seeds had the highest diversity, which could be attributed to the fact that these seeds were sampled within SERPs 1 – 10 or 20. However, the median diversity of Micro-collection seed was approximately the same as Expert seeds.

CHAPTER 10

CONTRIBUTIONS, FUTURE WORK, AND CONCLUSIONS

In order to memorialize an important story or event before it is lost due to link rot and content drift, curators hand-select seed URIs of news articles, images, videos etc. to be preserved in Web archive collections. For example, two months after the *2014 Western African Ebola Outbreak* was declared a Public Health Emergency of International Concern, an archivist at the National Library of Medicine collected seed URIs to be preserved. Similarly, two years after the *Flint Water Crisis*, archivists at Michigan State University collected seeds to be preserved for the Flint story. More recently, just as they have done on multiple occasions, the Internet Archive requested for social media users to contribute seeds for the *2020 Coronavirus Pandemic*. In spite of these efforts, there are multiple important stories such as the *2018 MSD High School Shooting*, for which we do not have archived collections due to a shortage of curators. The shortage problem which results in gaps in Web archive collections is further exacerbated by the lack of domain knowledge which is required to build these collections. Faced with the problem of a lack of Web archive collections for important stories and events, in this work, we asked if the seed generation process to bootstrap Web archive collections could be automated.

We addressed automating the seed generation process by exploiting the collective domain expertise of social media users by leveraging the Micro-collections they create. We formulated the three research questions to address the primary task of automating the seed generation process. We begin this final chapter with a review of the research questions and description of how they were addressed. Next, we state the contributions as a result of this research, outline areas of future research, and finally conclude.

- **RESEARCH QUESTION 1: How do we identify, extract, and profile Micro-collections in social media?**
- **RESEARCH QUESTION 2: Do seeds from Micro-collections differ from seeds from SERPs and hashtags?**

In Chapter 5, we addressed the first research question by introducing the novel *post class* system (Chapter 5.1) of labeling social media posts irrespective of platform. The ability to

identify social media posts across different platforms facilitates studying them without the concern of the cosmetic or operational differences of their parent social media platforms. For example, $\mathbf{P}_1\mathbf{A}_1$ posts map to Reddit or Twitter posts authored by a single user and visible from the SERP. The post class enabled the identification of Micro-collections (Chapter 5) as threaded conversations of social media posts created by single ($\mathbf{P}_1\mathbf{A}_n$) or multiple users ($\mathbf{P}_n\mathbf{A}_n$). Following the identification of Micro-collections, we profiled (Chapter 5.3) them by studying the distribution of their URIs, probability estimates for finding k seeds in a Micro-collection, and quantifying the precision/age of their URIs. The second research question was similarly addressed in Chapter 5 by an experiment conducted to compare seeds from Micro-collections ($\mathbf{P}_1\mathbf{A}_n \cup \mathbf{P}_n\mathbf{A}_n$) and seeds from SERPs ($\mathbf{P}_1\mathbf{A}_1$) for text and hashtag queries.

- **RESEARCH QUESTION 3: How do we evaluate automatically created collections with those generated by human experts in Archive-It?**

In Chapter 6, we made the first attempt to address the third research question by introducing of the Collection Characterizing Suite (CCS - Chapter 6.1), a suite of seven measures, for describing individual collections. The CCS also provides a means of comparing multiple collections (Chapter 6.1) by measuring the distances between their respective CCS vectors. This method was applied to check if collections generated automatically and semi-automatically from social media sources such as Storify, Reddit, Twitter, and Wikipedia were similar to Archive-It human-generated collections. The results showed that social media sources produce collections that are similar to Archive-It collections.

Additional contributions were made to the third research question in Chapter 7 with the introduction of Quality Proxies (QPs) for quantifying the quality of seeds. While Chapter 6 focused on comparison, Chapter 7 addressed the third research question by extending the comparison idea to evaluation by approximating the quality of the seeds across multiple dimensions such as *popularity* (7.2), *geographical* (Chapter 7.3.1), *temporal* (Chapter 7.3.2), *subject expert* (Chapter 7.3.3), *retrievability* (Chapter 7.3.4), *relevance* (Chapter 7.3.6), *reputation* (Chapter 7.3.5), and *scarcity* (Chapter 7.3.7). In Chapter 9, we presented our MCQP framework for bootstrapping Web archive collections from Micro-collections in social media. Our framework can be used generate seeds to augment existing Web archive collections such as the NLM *Ebola Virus* [3] and MSU *Flint Water Crisis* [33] collections, or create new archived collections for stories and events such as the *2014 MSD High School shooting*. We applied the Quality Proxies to assess seeds generated from Micro-collections

and human experts by quantifying the precision of the Micro-collection seeds selected by Quality Proxies when novelty is (Chapter 9.2.2) or is not (Chapter 9.3.1) prioritized.

10.1 CONTRIBUTIONS

Collection building encompasses seed selection, collection building with or without focused crawlers, and collection comparison. The major focuses of this research addressed the automation of the seed generation process. Seed selection research is sparse, and in addition to making contributions to the seed selection aspect of collection building, we made contributions to collection comparison as well. The following is an enumeration of the contributions of this work categorized according to their direct or indirect relationship to the primary concerns of this research:

1. Directly related

- (a) Chapter 5: We introduced a novel source for generating seeds from URIs in the threaded conversations of social media posts created by single or multiple users called Micro-collections.
- (b) Chapter 5: We provided the post class vocabulary, for labeling social media posts across different platforms.
- (c) Chapter 9: The first and second contributions culminated in the introduction of the MCQP framework for bootstrapping Web archive collections from Micro-collections in Social Media.
- (d) Chapter 7: We introduced the multi-dimensional Quality Proxies for seeds that express the individual quality trait of a seed within a single dimension (e.g., popularity or geographical proximity).
- (e) Chapter 8: We showed that different combinations of QPs map to different policies (e.g., prioritizing of popularity - rp , lk , sh or narrow reputation - re_n) fulfilling different seed selection goals, illustrating the versatility in seed selection the Quality Proxies offer.
- (f) Chapter 6 and Chapter 9.2: We demonstrated how to characterize/compare individual seeds (with Quality Proxies) and collections (with Quality Proxies and/or CCS)
- (g) Chapter 6, Chapter 7.5, and Chapter 9.2: We conducted multiple studies to characterize seeds generated from social media (e.g., Twitter, Reddit, Scoop.it)

and seeds generated from SERPs and/or experts on Archive-It. The results of these studies inform policies for generating seeds from SERPs and social media.

2. Indirectly related

- (a) We demonstrated the utility of SERPs beyond search by showing that SERPs may be used in classifying queries into categories such as *scholar* or *non-scholar* [140].
- (b) We quantified the well-known phenomena of the disappearance of news stories on the Google SERP and emphasized the need for collection building from SERPs to begin early and persist [24].

3. Software/Datasets/Services

- (a) Local Memory Project [197]: Suite of tools to build, archive and share collections of local news stories from local news sources
- (b) US and Non-US Local News Repository [198]: The US repository consists meta-data (website, Twitter handles, etc) of 5,992 Newspapers, 1,061 TV stations, and 2,539 Radio stations. The Non-US local news repository consists of 6,638 Newspapers from 183 countries and 3,151 cities.
- (c) Sungram [160]: A tool that summarizes a collection of text documents by generating the most frequent sumgrams (conjoined ngrams).
- (d) StoryGraph [199, 200, 201, 202]: A collection of tools that analyze the news cycle by computing the similarity of news stories across 17 US news sources.

10.2 FUTURE WORK

Thus far, we have applied our framework to extract seeds for topics affecting mostly English-speaking regions. A future research would investigate the extent to which our framework can be applied to non-English speaking regions. This poses some technical and semantic difficulties because many NLP tools are not language-agnostic and mostly target the English language. In Chapter 9.4, we presented some limitations of our framework and Quality Proxies. Future research effort would address these limitations. For example, improving the relevance evaluation to reduce false positive errors resulting from insufficient text and a high relevance threshold. Additionally, a future research would identify new

Quality Proxies and profile different combinations of them to highlight the properties of seeds they surface.

10.3 CONCLUSIONS

The Web is one of the greatest outcomes of human endeavor, but it has some major flaws, one of which is, the Web forgets. Web archive collections provide a crucial means of reducing the costly effects of link rot which causes the Web resources that chronicle important stories and events to disappear. These archived collections begin with seeds hand-selected by experts, selected by social media users in response to seed crowd-sourcing calls, or scraped from SERPs. Each of these methods for generating seeds is vital but insufficient in themselves, resulting in shortages of Web archive collections for many important stories and events. Research into seed selection for Web archive collections is sparse, consequently, this research effort explored the state of the art in seed selection, collection building, and collection comparison. We made contributions to seed selection by introducing a new source for seeds, Micro-collections - social media posts from the threaded conversation of single or multiple users. We studied and profiled Micro-collections, and seeds generated from SERPs. We made additional contributions to the quality assessment and comparison of seeds/collection through the introduction of the multi-dimensional Quality Proxies for seeds. The Quality Proxies assigns a quality trait of seed within a single dimension. Seeds can be assigned a quality score by selecting different combinations of Quality Proxies which map to different seed selection policies. We presented the MCQP framework for bootstrapping Web archive collections from Micro-collections and evaluated the framework. Our results showed that Quality Proxies resulted in the selection of quality seeds when novelty is and is not prioritized. We believe these contributions further the understanding of the seed selection, collection building, and collection evaluation.

REFERENCES

- [1] Centers for Disease Control and Prevention (CDC), “Years of Ebola Virus Disease Outbreaks.” <https://www.cdc.gov/vhf/ebola/history/chronology.html>, 2018.
- [2] Bell, BP and Damon, IK and Jernigan, DB and Kenyon, TA and Nichol, ST and O’Connor, JP and Tappero, JW, “Overview, Control Strategies, and Lessons Learned in the CDC Response to the 2014–2016 Ebola Epidemic.,” *Morbidity and Mortality Weekly Report (MMWR) Supplement*, vol. 65, no. 3, pp. 4–11, 2016. DOI: <http://dx.doi.org/10.15585/mmwr.su6503a2>.
- [3] National Library of Medicine, “Global Health Events.” <https://archive-it.org/collections/4887>, 2014.
- [4] Robert F. Worth, “How a Single Match Can Ignite a Revolution.” <https://www.nytimes.com/2011/01/23/weekinreview/23worth.html>, 2011.
- [5] NPR, “The Arab Spring: A Year Of Revolution.” <https://www.npr.org/2011/12/17/143897126/the-arab-spring-a-year-of-revolution>, 2011.
- [6] History.com, “Arab Spring.” <https://www.history.com/topics/middle-east/arab-spring>, 2018.
- [7] BBC, “Arab Uprisings.” <https://www.bbc.com/news/world-middle-east-12813859>, 2014.
- [8] Caryle Murphy, “The Arab Spring: The Uprising and its Significance.” <https://www.trinitydc.edu/magazine-2012/the-arab-spring-the-uprising-and-its-significance/>, 2012.
- [9] Amnesty International, “The ‘Arab Spring’: Five Years On.” <https://www.amnesty.org/en/latest/campaigns/2016/01/arab-spring-five-years-on/>, 2016.
- [10] Joseph V. Micallef, “The Arab Spring: Six Years Later.” https://www.huffingtonpost.com/joseph-v-micallef/the-arab-spring-six-years_b_14461896.html, 2018.

- [11] H. M. SalahEldeen and M. L. Nelson, “Losing my revolution: How many resources shared on social media have been lost?,” in *International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, pp. 125–137, 2012.
- [12] M. Klein, H. Shankar, and H. Van de Sompel, “Robust links in scholarly communication,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2018)*, pp. 357–358, 2018.
- [13] S. M. Jones, H. Van de Sompel, H. Shankar, M. Klein, R. Tobin, and C. Grover, “Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content,” *PloS one*, vol. 11, no. 12, 2016.
- [14] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin, “Scholarly context not found: one in five articles suffers from reference rot,” *PloS one*, vol. 9, no. 12, p. e115253, 2014.
- [15] J. Zittrain, K. Albert, and L. Lessig, “Perma: Scoping and addressing the problem of link and reference rot in legal citations,” *Legal Information Management*, vol. 14, p. 8899, 2014.
- [16] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, “Sic transit gloria telae: towards an understanding of the web’s decay,” in *Proceedings of the International Conference on World Wide Web (WWW 2014)*, pp. 328–337, ACM, 2004.
- [17] Tim Cushing, “Federal Judge Says Internet Archive’s Wayback Machine A Perfectly Legitimate Source Of Evidence.” <https://www.techdirt.com/articles/20160518/08175934474/federal-judge-says-internet-archives-wayback-machine-perfectly-legitimate-source-evidence.shtml>, 2016.
- [18] United States Court Of Appeals for the Second Circuit, “United States v. Gasperini, No. 17-2479 (2d Cir. 2018).” <https://law.justia.com/cases/federal/appellate-courts/ca2/17-2479/17-2479-2018-07-02.html>, 2018.
- [19] United States Agency for International Development (USAID), “On the Front Lines of an Epidemic: The Battle Against Ebola.” <http://wayback.archive-it.org/4887/20141022093244/http://blog.usaid.gov/ebola/>, 2014.

- [20] Centers for Disease Control and Prevention (CDC), “Thoughts from CDC Director Tom Frieden, MD, MPH.” <http://wayback.archive-it.org/4887/20141015093501/http://blogs.cdc.gov/cdcdirector/>, 2014.
- [21] Internet Archive Global Events, “2016 Pulse Nightclub Shooting Web Archive.” <http://archive-it.org/collections/7570>, 2016.
- [22] Internet Archive, “Help build an archive documenting responses to the 2016 U.S. presidential election at.” <https://twitter.com/internetarchive/status/797263535994613761>, 2016.
- [23] Internet Archive, “What web pages should we save concerning DAPL? Tell us here:” <https://twitter.com/internetarchive/status/806228431474028544>, 2016.
- [24] A. C. Nwala, M. C. Weigle, and M. L. Nelson, “Scraping SERPs for archival seeds: it matters when you start,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2018)*, pp. 263–272, 2018.
- [25] M. Klein, L. Balakireva, and H. Van de Sompel, “Focused Crawl of Web Archives to Build Event Collections,” in *Proceedings of the International ACM Web Science Conference (WebSci 2018)*, 2018.
- [26] Sean Rossman, “‘We’re children. You guys are the adults’: Shooting survivor, 17, calls out lawmakers.” <https://www.usatoday.com/story/news/nation-now/2018/02/15/were-children-you-guys-adults-shooting-survivor-17-calls-out-lawmakers/341002002/>, 2018.
- [27] Stephanie Ebbs, “Survivors of Florida high school shooting call for action on gun control.” <https://abcnews.go.com/Politics/survivors-florida-high-school-shooting-call-action-gun/story?id=53111278>, 2018.
- [28] Colin Dwyer, Camila Domonoske, and Emily Sullivan, “Survivors of Florida high school shooting call for action on gun control.” <https://www.npr.org/sections/tetwo-way/2018/02/28/589436112/dicks-sporting-goods-ends-sale-of-assault-style-rifles-citing-florida-shooting>, 2018.
- [29] Marwa Eltagouri, “Dick’s and Walmart raised the age for gun purchases. This 20-year-old is suing.” <https://www.washingtonpost.com/news/business/wp/2018/0>

- 3/06/a-20-year-old-is-suing-dicks-and-walmart-over-new-gun-policies-alleging-age-discrimination/, 2018.
- [30] German Lopez, “It’s official: March for Our Lives was one of the biggest youth protests since the Vietnam War.” <https://www.vox.com/policy-and-politics/2018/3/26/17160646/march-for-our-lives-crowd-size-count>, 2018.
- [31] D. Robbins, “ANALYSIS: How Michigan And National Reporters Covered The Flint Water Crisis.” <https://mediamatters.org/research/2016/02/02/analysis-how-michigan-and-national-reporters-co/208290>, 2016.
- [32] R. Fonger, “State says Flint River water meets all standards but more than twice the hardness of lake water.” http://www.mlive.com/news/flint/index.ssf/2014/05/state_says_flint_river_water_m.html, 2014.
- [33] Michigan State University, “Flint Water Crisis Websites Archive.” <https://archive-it.org/collections/6811>, 2016.
- [34] Doing Things Differently, “Tweet.” <https://twitter.com/dtdchange/status/676902666153406465>, 2015.
- [35] Wikipedia, “Stoneman Douglas High School shooting.” https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting, 2018.
- [36] S. M. Jones, “Where Can We Post Stories Summarizing Web Archive Collections?.” <http://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html>, 2017.
- [37] Twitter Moments, “17 people are dead after school shooting in Florida.” <https://twitter.com/i/moments/963863619271254016>, 2018.
- [38] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM (JACM 1999)*, vol. 46, no. 5, pp. 604–632, 1999.
- [39] Jesse O’Riordan, “Protests In Kiev Turn Violent.” <https://storify.com/jesseoriordan/ukraine-fight-for-their-rights>, 2014.
- [40] Internet Archive Global Events, “Ukraine Conflict.” <https://archive-it.org/collections/4399/>, 2014.

- [41] S. M. Jones, A. Nwala, M. C. Weigle, and M. L. Nelson, “The Many Shapes of Archive-It,” in *Proceedings of the 15th International Conference on Digital Preservation (iPres 2018)*, 2018.
- [42] T. Berners-Lee, “The original proposal of the WWW.” <https://www.w3.org/History/1989/proposal.html>, 1990.
- [43] Berners-Lee, Tim and Bray, Tim and Connolly, Dan and Cotton, Paul and Fielding, Roy and Jeckle, Mario and Lilley, Chris and Mendelsohn, Noah and Orchard, David and Walsh, Norman and Williams, Stuart, “Architecture of the World Wide Web, Volume One.” <https://www.w3.org/TR/webarch/#def-world-wide-web>, 2002.
- [44] R. e. Lewis, “Dereferencing HTTP URIs.” <https://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14>, 2007.
- [45] I. Hickson, R. Berjon, S. Faulkner, T. Leithead, E. D. Navara, E. O’Connor, and S. Pfeiffer, “HTML 5.2 W3C Recommendation.” <http://web.archive.org/web/20170319034643/www.w3.org/TR/html5/>, 2017.
- [46] P. J. Leach, T. Berners-Lee, J. C. Mogul, L. Masinter, R. T. Fielding, and J. Gettys, “RFC 2616.” <https://tools.ietf.org/html/rfc2616.html>, 1999.
- [47] J. Cho, H. Garcia-Molina, and L. Page, “Efficient crawling through url ordering,” *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 161–172, 1998.
- [48] M. H. Alam, J. Ha, and S. Lee, “Novel approaches to crawling important pages early,” *Knowledge and Information Systems*, vol. 33, no. 3, pp. 707–734, 2012.
- [49] M. Najork and J. L. Wiener, “Breadth-first crawling yields high-quality pages,” in *Proceedings of the International Conference on World Wide Web (WWW 2001)*, pp. 114–118, 2001.
- [50] C. Castillo, M. Marin, A. Rodriguez, and R. Baeza-Yates, “Scheduling algorithms for web crawling,” in *WebMedia and LA-Web, 2004. Proceedings*, pp. 10–17, 2004.
- [51] S. M. Mirtaheri, M. E. Dingtürk, S. Hooshmand, G. V. Bochmann, G.-V. Jourdan, and I. V. Onut, “A Brief History of Web Crawlers,” in *Conference of the Center for Advanced Studies on Collaborative Research (CASCON 2013)*, pp. 40–54, 2013.

- [52] T. Seymour, D. Frantsvog, and S. Kumar, "History of search engines," *International Journal of Management & Information Systems (IJMIS 2011)*, vol. 15, no. 4, pp. 47–58, 2011.
- [53] Aaron Wall, "Search Engine History." <http://www.searchenginehistory.com/>, 2018.
- [54] Matthew Gray, "Internet Growth and Statistics: Credits and Background." <http://www.mit.edu/~mkgray/net/background.html>, 2018.
- [55] D. Eichmann, "The RBSE Spider - Balancing Effective Search Against Web Load," in *Proceedings of the International Conference on World Wide Web (WWW 1994)*, 1994.
- [56] O. A. McBryan, "GENVL and WWW: Tools for Taming the Web," in *Proceedings of the International Conference on World Wide Web (WWW 1994)*, 1994.
- [57] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [58] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," *Computer Networks and ISDN systems*, vol. 30, no. 1-7, pp. 65–74, 1998.
- [59] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999.
- [60] I. Ben-Shaul, M. Herscovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, V. Soroka, and S. Ur, "Adding support for dynamic and focused search with fetuccino," *Computer Networks*, vol. 31, no. 11-16, pp. 1653–1665, 1999.
- [61] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pp. 104–111, 1998.
- [62] S. J. Carrière and R. Kazman, "Webquery: Searching and visualizing the web through connectivity," *Computer Networks and ISDN Systems*, vol. 29, no. 8, pp. 1257–1268, 1997.

- [63] B. D. Davison, "Topical locality in the web," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pp. 272–279, 2000.
- [64] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs.," in *Proceedings of Very Large Data Bases (VLDB 2000)*, pp. 527–534, 2000.
- [65] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring web communities from link topology," in *Proceedings of ACM Hypertext and Hypermedia (HT 1998)*, pp. 225–234, 1998.
- [66] V. Kluev, "Compiling document collections from the Internet," in *ACM SIGIR Forum*, vol. 34, pp. 9–14, 2000.
- [67] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (salsa) and the tkc effect1," *Computer Networks*, vol. 33, no. 1-6, pp. 387–401, 2000.
- [68] S. Chakrabarti, B. Dom, D. Gibson, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Experiments in topic distillation," in *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, Melbourne, Australia, 1998.
- [69] J. Dean and M. R. Henzinger, "Finding related pages in the World Wide Web," *Computer networks*, vol. 31, no. 11-16, pp. 1467–1479, 1999.
- [70] E. Garfield, *Mapping the Structure of Science*. Wiley & Sons, 1979.
- [71] P. Mutschke, "Enhancing information retrieval in federated bibliographic data sources using author network based stratagems," in *International Conference on Theory and Practice of Digital Libraries (TPDL 2001)*, pp. 287–299, Springer, 2001.
- [72] G. Salton, *Automatic information organization and retrieval*. McGraw Hill Text, 1968.
- [73] M. Klein, M. L. Nelson, and J. Z. Pao, "Augmenting OAI-PMH repository holdings using search engine APIs," in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JC DL 2007)*, pp. 486–486, 2007.
- [74] M. Klein, M. L. Nelson, and J. Z. Pao, "OAI-PMH Repository Enhancement for the NASA Langley Research Center Atmospheric Sciences Data Center," in *Proceedings of International Web Archiving Workshop (IWA W 2007)*, 2007.

- [75] Y. Liu and A. Agah, “Topical Crawling on the Web through Local Site-Searches,” *Journal of Web Engineering (JWE 2013)*, vol. 12, no. 3&4, pp. 203–214, 2013.
- [76] M. Klein and M. L. Nelson, “Moved but not gone: an evaluation of real-time methods for discovering replacement Web pages,” *International Journal on Digital Libraries (IJDL 2014)*, vol. 14, no. 1-2, pp. 17–38, 2014.
- [77] F. McCown, J. A. Smith, and M. L. Nelson, “Lazy preservation: Reconstructing websites by crawling the crawlers,” in *Proceedings of the International workshop on Web Information and Data Management (WIDM 2006)*, pp. 67–74, 2006.
- [78] F. McCown and M. L. Nelson, “Agreeing to disagree: search engines and their public interfaces,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007)*, pp. 309–318, 2007.
- [79] F. McCown and M. L. Nelson, “Search Engines and Their Public Interfaces: Which APIs are the Most Synchronized?,” in *Proceedings of the International Conference on World Wide Web (WWW 2007)*, pp. 1197–1198, 2007.
- [80] H. Hockx-Yu, “The past issue of the Web,” in *Proceedings of the International ACM Web Science Conference (WebSci 2011)*, pp. 1–8, 2011.
- [81] M. Burner and B. Kahle, “Arc file format.” <https://archive.org/web/researcher/ArcFileFormat.php>, 1996.
- [82] ISO, “ISO 28500:2009: Information and documentation WARC file format,” 2009.
- [83] D. Gomes, J. Miranda, and M. Costa, “A survey on web archiving initiatives,” in *International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*, pp. 408–420, 2011.
- [84] J. Bailey, A. Grotke, E. McCain, C. Moffatt, and N. Taylor, “Web Archiving in the United States: A 2016 Survey.” https://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf, 2016.
- [85] H. Van de Sompel, M. L. Nelson, and R. Sanderson, “HTTP Framework for Time-Based Access to Resource States – Memento, Internet RFC 7089.” <https://tools.ietf.org/html/rfc7089>, 2013.

- [86] Memento, “Memento Guide - Introduction to Memento.” <http://www.mementoweb.org/guide/quick-intro/>, 2015.
- [87] Progressive Turnout Project (@TurnoutPAC), “Tweet.” <https://twitter.com/TurnoutPAC/status/1078060603036585990>, 2018.
- [88] LordVelaryon, “Reddit Post.” https://www.reddit.com/r/soccer/comments/8wu97n/next_day_discussion_postmatch_thread_brazil_12/, 2018.
- [89] Storify, “Storify End-of-Life.” <http://web.archive.org/web/20190411150221/https://storify.com/faq-eol>, 2018.
- [90] Y. AlNoamany, M. C. Weigle, and M. L. Nelson, “Characteristics of social media stories,” in *International Conference on Theory and Practice of Digital Libraries (TPDL 2015)*, pp. 267–279, 2015.
- [91] Wikipedia, “EWikipedia:Statistics.” <https://en.wikipedia.org/wiki/Wikipedia:Statistics>, 2018.
- [92] Aliza Rosen, “Tweeting Made Easier.” https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html, 2017.
- [93] Madhu Muthukumar, “Moments, the best of Twitter in an instant.” https://blog.twitter.com/official/en_us/a/2015/moments-the-best-of-twitter-in-an-instant-0.html, 2015.
- [94] Drew Olanoff, “Twitter Debuts Moments.” <https://techcrunch.com/2015/10/06/project-glacier/>, 2015.
- [95] Tammy Duckworth, “Tweet.” <https://twitter.com/SenDuckworth/status/832710593207431171>, 2017.
- [96] March for Science, “Tweet.” <https://twitter.com/ScienceMarchDC/status/847504543604752385>, 2017.
- [97] Sasank Reddy, “Nice Threads.” https://blog.twitter.com/official/en_us/topics/product/2017/nicethreads.html, 2015.
- [98] Sarah Perez, “Twitter officially launches threads, a new feature for easily posting tweetstorms.” <https://techcrunch.com/2017/12/12/twitter-officially-launches-threads-a-new-feature-for-easily-writing-tweetstorms/>, 2017.

- [99] Asemeyibo Buowari-Brown, “The Battle Raging In Nigeria Over Control Of Oil — VICE on HBO.” <https://www.facebook.com/dagogo.buowaribrown/posts/10156118605352295>, 2018.
- [100] VICE News, “The Battle Raging In Nigeria Over Control Of Oil — VICE on HBO.” https://youtu.be/vAgw_Zyznx0, 2018.
- [101] jazir5, “Reddit Post Comment.” https://www.reddit.com/r/science/comments/7k88s4/ebola_survivors_still_immune_to_virus_after_40/drctlx1/, 2017.
- [102] D. Bergmark, “Collection synthesis,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, pp. 253–262, 2002.
- [103] A. Heydon and M. Najork, “Mercator: A scalable, extensible web crawler,” *Proceedings of the International Conference on World Wide Web (WWW 1999)*, vol. 2, no. 4, pp. 219–229, 1999.
- [104] M. M. Farag, S. Lee, and E. A. Fox, “Focused crawler for events,” *International Journal on Digital Libraries (IJDL)*, vol. 19, no. 1, pp. 3–19, 2018.
- [105] T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavarakas, and P. Senellart, “Exploiting the social and semantic web for guided web archiving,” in *International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, pp. 426–432, 2012.
- [106] T. Risse, E. Demidova, S. Dietze, W. Peters, N. Papailiou, K. Doka, Y. Stavarakas, V. Plachouras, P. Senellart, F. Carpentier, *et al.*, “The ARCOMEM architecture for social-and semantic-driven web archiving,” *Future Internet*, vol. 6, no. 4, pp. 688–716, 2014.
- [107] G. Gossen, E. Demidova, and T. Risse, “iCrawl: improving the freshness of web collections by integrating social web and focused web crawling,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2015)*, pp. 75–84, 2015.
- [108] G. Gossen, E. Demidova, and T. Risse, “Analyzing web archives through topic and event focused sub-collections,” in *Proceedings of the International ACM Web Science Conference (WebSci 2016)*, pp. 291–295, 2016.
- [109] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge,” in *Proceedings*

- of the *SIGMOD International Conference on Management of Data (SIGMOD 2008)*, pp. 1247–1250, 2008.
- [110] G. Gossen, E. Demidova, and T. Risse, “Extracting event-centric document collections from large-scale web archives,” in *International Conference on Theory and Practice of Digital Libraries (TPDL 2017)*, pp. 116–127, 2017.
- [111] F. Nanni, S. P. Ponzetto, and L. Dietz, “Building entity-centric event collections,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2017)*, pp. 199–208, 2017.
- [112] F. Nanni, S. P. Ponzetto, and L. Dietz, “Toward comprehensive event collections,” *International Journal on Digital Libraries (IJDL 2018)*, vol. 21, no. 2, pp. 215–229, 2018.
- [113] P. Ferragina and U. Scaiella, “Tagme: on-the-fly annotation of short text fragments (by wikipedia entities),” in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2010)*, pp. 1625–1628, 2010.
- [114] J. Dalton, L. Dietz, and J. Allan, “Entity query feature expansion using knowledge base links,” in *Proceedings of ACM SIGIR conference on Research and development in Information Retrieval (SIGIR 2014)*, pp. 365–374, 2014.
- [115] X. Liu and H. Fang, “Latent entity space: a novel retrieval approach for entity-bearing queries,” *Information Retrieval Journal*, vol. 18, no. 6, pp. 473–503, 2015.
- [116] P. Ristoski and H. Paulheim, “RDF2Vec: RDF graph embeddings for data mining,” in *International Semantic Web Conference (ISWC 2016)*, pp. 498–514, 2016.
- [117] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1532–1543, 2014.
- [118] Wikipedia, “Wikipedia:Wikipedians.” <https://en.wikipedia.org/wiki/Wikipedia:Wikipedians>, 2018.
- [119] S. Zheng, P. Dmitriev, and C. L. Giles, “Graph based crawler seed selection,” in *Proceedings of the International Conference on World Wide Web (WWW 2009)*, pp. 1089–1090, 2009.

- [120] R. Prasath and P. Öztürk, “Finding potential seeds through rank aggregation of web searches,” in *International Conference on Pattern Recognition and Machine Intelligence (ICPRAI 2011)*, pp. 227–234, 2011.
- [121] Y. Du, Y. Hai, C. Xie, and X. Wang, “An approach for selecting seed urls of focused crawler based on user-interest ontology,” *Applied Soft Computing*, vol. 14, pp. 663–676, 2014.
- [122] S. Yang, K. Chitturi, G. Wilson, M. Magdy, and E. A. Fox, “A study of automation from seed url generation to focused web archive development: the ctnet context,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2012)*, pp. 341–342, 2012.
- [123] P. N. Priyatam, A. Dubey, K. Perumal, S. Praneeth, D. Kakadia, and V. Varma, “Seed selection for domain-specific search,” in *Proceedings of the International Conference on World Wide Web (WWW 2014)*, pp. 923–928, 2014.
- [124] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, “CiteSeerX: an architecture and web service design for an academic document search engine,” in *Proceedings of the International Conference on World Wide Web (WWW 2006)*, pp. 883–884, ACM, 2006.
- [125] Russell Goldman and Alyssa Newcomb, “Elementary School Clerk Says She Convinced Suspect to Put His Weapons Down and Surrender: Exclusive.” <https://abcnews.go.com/US/elementary-school-clerk-convinced-suspect-put-weapons-surrender/story?id=20014879>, 2013.
- [126] Greg Botelho, Vivian Kuo, and Josh Levs, “Antoinette Tuff hailed as ‘true hero’ for handling Georgia school gunman.” <https://www.cnn.com/2013/08/21/us/georgia-school-gunshots/index.html>, 2013.
- [127] G. Gossen, E. Demidova, and T. Risse, “The iCrawl wizard—supporting interactive focused crawl specification,” in *European Conference on Information Retrieval (ECIR 2015)*, pp. 797–800, Springer, 2015.
- [128] G. S. Bonn, “Evaluation of the collection,” *Library Trends*, vol. 22, no. 3, pp. 265–304, 1974.
- [129] L. Carnovsky, “A list of books for college libraries,” *The Library Quarterly: Information, Community, Policy*, vol. 2, no. 2, pp. 161–164, 1932.

- [130] N. E. Gwinn and P. H. Mosher, "Coordinating collection development: The RLG conspectus," *College & Research Libraries*, vol. 44, no. 2, pp. 128–140, 1983.
- [131] A. W. Ferguson, J. Grant, and J. S. Rutstein, "The RLG conspectus: its uses and benefits," *College & Research Libraries*, vol. 49, no. 3, pp. 197–206, 1988.
- [132] D. Lesniaski, "Evaluating collections: a discussion and extension of brief tests of collection strength," *College & Undergraduate Libraries*, vol. 11, no. 1, pp. 11–24, 2004.
- [133] H. D. White, *Brief tests of collection strength: A methodology for all types of libraries*. No. 88, Greenwood Publishing Group, 1995.
- [134] B. Lockett, *Guide to the evaluation of library collections*. American Library Association, 1989.
- [135] T. Heidenwolf, "Evaluating an interdisciplinary research collection," *Collection Management*, vol. 18, no. 3-4, pp. 33–48, 1994.
- [136] A. C. Nwala, M. C. Weigle, and M. L. Nelson, "Bootstrapping web archive collections from social media," in *Proceedings of ACM Hypertext and Social Media (HT 2018)*, pp. 64–72, 2018.
- [137] T. Risse, E. Demidova, and G. Gossen, "What Do You Want to Collect from the Web?," in *Proceedings of Building Web Observatories Workshop (BWOW 2014)*, 2014.
- [138] A. Broder, "A Taxonomy of Web Search," in *ACM SIGIR forum*, vol. 36, pp. 3–10, 2002.
- [139] J. Kim and V. R. Carvalho, "An analysis of time-instability in web search results," in *European Conference on Information Retrieval (ECIR 2011)*, pp. 466–478, 2011.
- [140] A. C. Nwala and M. L. Nelson, "A supervised learning algorithm for binary domain classification of web queries using SERPs," in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2016)*, pp. 237–238, 2016.
- [141] A. C. Nwala, "Scraping SERPs for archival seeds: it matters when you start - Git Repo." <https://github.com/anwala/SERPRefind>, 2018.

- [142] A. C. Nwala, M. C. Weigle, A. B. Ziegler, A. Aizman, and M. L. Nelson, “Local memory project: Providing tools to build collections of stories for local events from local sources,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2017)*, pp. 219–228, 2017.
- [143] A. C. Nwala, “Local memory project - local stories collection generator.” <https://chrome.google.com/webstore/detail/local-memory-project/khineeknpgfcholchjihimhofilcfp>, 2016.
- [144] J. F. Brunelle, M. C. Weigle, and M. L. Nelson, “Archiving deferred representations using a two-tiered crawling approach,” *Proceedings of iPRES 2015*, 2015.
- [145] A. C. Nwala, M. C. Weigle, and M. L. Nelson, “Using Micro-collections in Social Media to Generate Seeds for Web Archive Collections,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2019)*, pp. 251–260, 2019.
- [146] Scott Gottlieb, MD, “Tweet.” <https://twitter.com/ScottGottliebMD/status/1291117316831432715>, 2020.
- [147] Ilsensine, “[Ebola] 2014 Outbreak Report.” https://www.reddit.com/r/OutbreakNews/comments/2cn9yq/ebola_2014_outbreak_report/, 2014.
- [148] Ilsensine, “[Ebola] 2014 Outbreak Report.” https://www.reddit.com/r/OutbreakNews/comments/2cn9yq/ebola_2014_outbreak_report/, 2014.
- [149] Doing Things Differently, “Twitter User.” <https://twitter.com/dtdchange/>, 2012.
- [150] Ilsensine, “Reddit User.” <https://www.reddit.com/user/Ilsensine>, 2013.
- [151] WHO, “Statement on the 1st meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa.” <https://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/>, 2014.
- [152] Wikipedia, “West African Ebola virus epidemic.” https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic, 2014.
- [153] Centers for Disease Control and Prevention (CDC), “Years of Ebola Virus Disease Outbreaks.” <https://www.cdc.gov/vhf/ebola/history/chronology.html>, 2019.
- [154] Wikipedia, “Flint Water Crisis.” https://en.wikipedia.org/wiki/Flint_water_crisis, 2016.

- [155] Wikipedia, “Stoneman Douglas High School shooting.” https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting, 2017.
- [156] Wikipedia, “2018 FIFA World Cup.” https://en.wikipedia.org/wiki/2018_FIFA_World_Cup, 2018.
- [157] Wikipedia, “2018 United States elections.” https://en.wikipedia.org/wiki/2018_United_States_elections, 2018.
- [158] A. C. Nwala, “Using Micro-collections in Social Media to Generate Seeds for Web Archive Collections - Git Repo.” <https://github.com/anwala/MicroCollections> JCDL2019, 2019.
- [159] H. M. SalahEldeen and M. L. Nelson, “Carbon dating the web: estimating the age of web resources,” Tech. Rep. arXiv:1304.5213, 2013.
- [160] A. C. Nwala, “Sumgram: a tool that summarizes a collection of text documents by generating the most frequent sumgrams (conjoined ngrams).” <https://github.com/oduwsdl/sumgram/>, 2019.
- [161] G. C. Atkins, A. C. Nwala, M. C. Weigle, and M. L. Nelson, “Measuring News Similarity Across Ten US News Sites,” Tech. Rep. arXiv:1806.09082, 2018.
- [162] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling,” in *Proceedings of Association for Computational Linguistics (ACL 2005)*, pp. 363–370, 2005.
- [163] A. X. Chang and C. D. Manning, “Sutime: A library for recognizing and normalizing time expressions,” in *Proceedings of Language Resources and Evaluation Conference (LREC 2012)*, pp. 3735–3740, 2012.
- [164] A. C. Nwala, “An exploration of URL diversity measures.” <http://ws-dl.blogspot.com/2018/05/2018-05-04-exploration-of-url-diversity.html>, 2018.
- [165] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, “How much of the web is archived?,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011)*, pp. 133–136, 2011.

- [166] S. Alam and M. L. Nelson, “MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2016)*, pp. 243–244, 2016.
- [167] A. C. Nwala, “Finding URLs on Twitter - A simple recommendation.” <http://ws-dl.blogspot.com/2017/01/2017-01-23-finding-urls-on-twitter.html>, 2017.
- [168] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, “Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel,” Tech. Rep. Research Branch Report 8-75, Naval Technical Training Command Millington TN Research Branch, 1975.
- [169] M. Coleman and T. L. Liau, “A computer readability formula designed for machine scoring.,” *Journal of Applied Psychology*, vol. 60, no. 2, pp. 283–284, 1975.
- [170] E. A. Smith and R. Senter, “Automated readability index,” Tech. Rep. AMRL-TR-66-220, AMRL-TR. Aerospace Medical Research Laboratories (US), 1967.
- [171] M. Lichman, “UCI machine learning repository.” <http://archive.ics.uci.edu/ml>, 2013.
- [172] Danny Sullivan, “Dear Bing, We Have 10,000 Ranking Signals To Your 1,000. Love, Google.” <https://searchengineland.com/bing-10000-ranking-signals-google-55473>, 2020.
- [173] S. C. Woolley and D. Guilbeault, “Computational propaganda in the United States of America: Manufacturing consensus online.” <https://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2017/06/Comprop-USA.pdf>, 2017.
- [174] A. Marwick and R. Lewis, “Media manipulation and disinformation online.” <https://datasociety.net/pubs/oh/DataAndSocietyMediaManipulationAndDisinformationOnline.pdf>, 2017.
- [175] Nauman Siddique, “2019-09-10: Twitter Follower Growth for the 2020 Democratic Candidates.” <https://ws-dl.blogspot.com/2019/09/2019-09-10-twitter-follower-growth-for.html>, 2019.
- [176] A. Gupta and P. Kumaraguru, “Credibility ranking of tweets during high impact events,” in *Proceedings of the Workshop on Privacy and Security in Online Social Media*, pp. 2–8, ACM, 2012.

- [177] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum, “An empirical study on learning to rank of tweets,” in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 295–303, Association for Computational Linguistics, 2010.
- [178] R. Nagmoti, A. Teredesai, and M. De Cock, “Ranking approaches for microblog search,” in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 153–157, IEEE Computer Society, 2010.
- [179] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna, “Poultry markets: on the underground economy of twitter followers,” *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 527–532, 2012.
- [180] C. Yang, R. C. Harkreader, and G. Gu, “Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers,” in *International Workshop on Recent Advances in Intrusion Detection*, pp. 318–337, Springer, 2011.
- [181] Google, “Place Search.” <https://developers.google.com/places/web-service/search>, 2020.
- [182] L. Azzopardi and V. Vinay, “Retrievability: an evaluation measure for higher order information access tasks,” in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2008)*, pp. 561–570, 2008.
- [183] M. C. Traub, T. Samar, J. Van Ossenbruggen, J. He, A. de Vries, and L. Hardman, “Querylog-based assessment of retrievability bias in a large newspaper corpus,” in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2016)*, pp. 7–16, 2016.
- [184] H. Ramadan and J. Shantz, *Manufacturing Phobias: The political production of fear in theory and practice*. University of Toronto Press, 2016.
- [185] Bill Chappell, “U.N. Chief Targets ‘Dangerous Epidemic Of Misinformation’ On Coronavirus.” <https://www.npr.org/sections/coronavirus-live-updates/2020/04/14/834287961/u-n-chief-targets-dangerous-epidemic-of-misinformation-on-coronavirus>, 2020.

- [186] Andy Carvin and Graham Brookie, “Here’s How to Fight Coronavirus Misinformation.” <https://www.theatlantic.com/ideas/archive/2020/03/heres-how-fight-coronavirus-misinformation/608914/>, 2020.
- [187] A. Mian and S. Khan, “Coronavirus: the spread of misinformation,” *BMC medicine*, vol. 18, no. 1, pp. 1–2, 2020. DOI: <https://doi.org/10.1186/s12916-020-01556-3>.
- [188] Alastair Reid, “The 6 types of coronavirus misinformation to watch out for.” <https://firstdraftnews.org/latest/the-6-types-of-coronavirus-misinformation-to-watch-out-for/>, 2020.
- [189] Travis M. Andrews, “Why dangerous conspiracy theories about the virus spread so fast and how they can be stopped.” <https://www.washingtonpost.com/technology/2020/05/01/5g-conspiracy-theory-coronavirus-misinformation/>, 2020.
- [190] Flora Carmichael and Marianna Spring, “Coronavirus: Here’s how you can stop bad information from going viral.” <https://www.bbc.com/news/blogs-trending-51967889>, 2020.
- [191] A. C. Nwala, “Bootstrapping Web Archive Collections From Micro-collections In Social Media - Git Repo.” <https://github.com/anwala/dissertation>, 2020.
- [192] Archive-It API 1.0, “Archive-It.” <https://partner.archive-it.org/api/>, 2020.
- [193] National Library of Medicine, “Global Health Events Web Archive - Coronavirus.” <https://archive-it.org/collections/4887?fc=websiteGroup%3ACoronavirus+disease+%28COVID-19%29+outbreak>, 2020.
- [194] Internet Archive Global Events, “Archive-It Hurricane Harvey 2017.” <https://archive-it.org/collections/9323>, 2017.
- [195] National Library of Medicine (NLM), “Global Health Events Web Archive - Ebola Virus.” <https://archive-it.org/collections/4887?fc=websiteGroup%3AEbola+Outbreak+2014>, 2014.
- [196] S. Alam, M. L. Nelson, H. Van de Sompel, L. L. Balakireva, H. Shankar, and D. S. Rosenthal, “Web archive profiling through CDX summarization,” *International Journal on Digital Libraries*, vol. 17, no. 3, pp. 223–238, 2016.

- [197] Local Memory Project, “Local Memory Project.” <http://www.localmemory.org/>, 2020.
- [198] A. C. Nwala, “Local Memory Project Source Code.” <https://github.com/harvard-lil/local-memory>, 2017.
- [199] A. C. Nwala, “StoryGraph: Live news similarity (story link detection) measurement.” <http://storygraph.cs.odu.edu/>, 2020.
- [200] A. C, M. C. Weigle, and M. L. Nelson, “365 dots in 2019: Quantifying attention of news sources,” Tech. Rep. arXiv:2003.09989, 2020.
- [201] A. C. Nwala, “365 dots in 2018 - top news stories of 2018.” <https://ws-dl.blogspot.com/2019/03/2019-03-05-365-dots-in-2018-top-news.html>, 2019.
- [202] A. C. Nwala, “365 dots in 2019 - top news stories of 2019.” <https://ws-dl.blogspot.com/2020/01/2020-01-04-365-dots-in-2019-top-news.html>, 2020.

APPENDIX A

**EVALUATION RESULTS: ADDITIONAL TABLES FOR TOP
10 OVERLAP (WITH P@10) FOR SEEDS SCORED BY 1 – 3
QP COMBINATIONS**

TABLE 46: (Chapter 9.3.1, 2018 World Cup): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap		P@10				QP combinations
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}	GM
1	.75	.20	.67	.50	.25	.25	$ge_a, re_b, \overline{re_n}$
2	.75	.19	.67	.80	.40	.29	$\overline{ge_a}, re_b, \overline{re_n}$
3	.60	.19	.75	.86	.78	.43	$\overline{rp}, re_b, \overline{re_n}$
4	.60	.18	.83	.86	.71	.14	$\overline{sh}, \overline{sc}, re_b$
5	.60	.18	.75	.71	.50	.38	$\overline{sh}, re_b, \overline{re_n}$
6	.60	.17	.83	.67	.78	.50	$\overline{lk}, \overline{sc}, re_b$
7	.60	.15	.75	.50	.67	.75	$\overline{lk}, re_b, \overline{re_n}$
8	.50	.15	.86	.57	.75	.83	\overline{sh}, re_b
9	.50	.15	.86	1.0	.78	.60	\overline{lk}, re_b
10	.50	.15	.86	.63	.67	.33	rp, \overline{sh}, re_b
Averages							
	.60	.17	.78	.71	.63	.45	

TABLE 47: (Chapter 9.3.1, Hurricane Harvey (collected 2020)): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap		P@10				QP combinations
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}	GM
1	.50	.19	.89	.13	.20	.00	rl, re_b
2	.50	.19	.78	.14	.00	.00	sc, re_b
3	.50	.13	.89	.33	.20	.20	rl, \overline{sc}, re_b
4	.50	.13	.89	.56	.20	.00	rl, re_b, re_n
5	.50	.13	.63	.22	.22	.00	\overline{rp}, sc, re_b
6	.50	.13	.63	.22	.22	.00	$\overline{ge_d}, re_b, \overline{re_n}$
7	.50	.13	.67	.44	.00	.14	sc, re_b, re_n
8	.50	.13	.71	.44	.22	.29	$sc, re_b, \overline{re_n}$
9	.50	.13	.78	.22	.00	.00	rt, re_b
10	.50	.13	.78	.13	.00	.00	rt, \overline{sc}, re_b
Averages							
	.50	.14	.77	.28	.13	.06	

TABLE 48: (Chapter 9.3.1, Hurricane Harvey (collected 2017)): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap				P@10								QP combinations	
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$	GM	EM
1	.71	.33	.31	.17	.89	.30	.80	.13	.60	.00	.13	.00	rt, re_b	\bar{sc}
2	.71	.33	.31	.17	1.0	.30	.89	.11	.60	.00	.13	.13	rl, rt, re_b	\bar{rt}, \bar{sc}
3	.71	.25	.31	.15	1.0	.33	.67	.20	.60	.00	.13	.00	\bar{sh}, rt, re_b	\bar{sc}, re_n
4	.71	.25	.29	.15	1.0	.33	.70	.11	.60	.00	.33	.00	\bar{lk}, rt, re_b	\bar{rt}, \bar{sc}, re_n
5	.71	.22	.29	.15	.89	.00	.88	.00	.60	.00	.25	.14	rt, \bar{sc}, re_b	\bar{rl}, \bar{rp}, re_b
6	.71	.22	.27	.15	.89	.00	.70	.22	.60	.00	.40	.11	rt, re_b, re_n	\bar{rl}, \bar{rp}, re_n
7	.67	.22	.25	.14	.78	.00	.60	.22	.50	.00	.22	.00	rl, \bar{rt}, re_b	\bar{rl}, \bar{sh}, re_b
8	.63	.22	.25	.14	1.0	.00	.50	.00	.60	.00	.13	.17	\bar{rp}, rt, re_b	\bar{rl}, \bar{sh}, re_n
9	.63	.22	.23	.14	.78	.00	.89	.11	.60	.00	.38	.00	lk, rt, re_b	\bar{rl}, \bar{lk}, re_b
10	.60	.22	.23	.13	.78	.00	.80	.22	.30	.00	.33	.00	re_n	\bar{rl}, \bar{lk}, re_n
Averages														
	.68	.25	.27	.15	.90	.13	.74	.13	.56	.00	.24	.06		

TABLE 49: (Chapter 9.3.1, Flint Water Crisis): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap		P@10				QP combinations
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}	GM
1	.50	.21	.89	.88	.86	.67	$ge_a, \overline{sc}, \overline{re}_b$
2	.50	.19	1.0	.71	.50	.56	rt, re_b
3	.50	.19	1.0	.78	.50	.20	rt, sc, re_b
4	.50	.19	1.0	.75	.50	.63	rt, \overline{sc}, re_b
5	.45	.18	.89	.88	.44	.63	$sh, \overline{sc}, \overline{re}_n$
6	.45	.18	.89	.86	.44	.75	$lk, \overline{sc}, \overline{re}_n$
7	.44	.18	1.0	.89	.40	.56	dp, ge_a
8	.44	.13	1.0	.90	.40	.57	dp, ge_a, \overline{rt}
9	.43	.13	.88	.78	1.0	.60	$\overline{rp}, rt, \overline{sc}$
10	.43	.13	.88	.50	1.0	.14	$\overline{sh}, rt, \overline{sc}$
Averages							
	.47	.17	.94	.79	.60	.53	

TABLE 50: (Chapter 9.3.1, 2014 Ebola Virus Outbreak): Top 10 overlap for seeds scored by 1 – 3 QP combinations and the precision at 10 (P@10) for the respective QP combination/overlap scores. These values were produced by scoring (Equation 15) all seeds with a specific QP combination (e.g. rl, re_b), and measuring the overlap between the 10 top seeds from reference Google (G) or Expert (E) seeds and Micro-collections (M), and additionally measuring the P@10 for G, E, and M. **Key:** For **Overlap**, GM/EM - overlap between Google/Expert and Micro-collection. For **QP combinations**, GM/EM - QP combination used to score G/E and M seeds. For **P@10**, G/E - P@10 of Google/Expert seeds with respect to Micro-collections. M_G/M_E : P@10 of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

#	Overlap				P@10								QP combinations	
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$	GM	EM
1	.25	1.0	.18	.17	.75	.70	.71	.22	.80	.50	.25	.10	rt, \overline{re}_n	\overline{sc}
2	.25	1.0	.18	.14	.71	.70	.89	.50	.44	.50	.13	.22	$\overline{lk}, rt, \overline{re}_n$	re_b
3	.25	1.0	.17	.14	.75	.67	.88	.50	.80	.22	.38	.11	$ge_a, rt, \overline{re}_n$	$\overline{rp}, \overline{sc}$
4	.25	1.0	.17	.13	.75	.67	1.0	1.0	.80	.22	.29	.20	$\overline{ge}_a, rt, \overline{re}_n$	\overline{rp}, re_b
5	.25	1.0	.17	.13	.75	.67	.63	.38	.80	.33	.38	.11	$rt, \overline{sc}, \overline{re}_n$	$\overline{sh}, \overline{sc}$
6	.22	1.0	.17	.13	.67	.67	.89	.67	.50	.33	.00	.11	rp, rt, \overline{re}_n	$\overline{lk}, \overline{sc}$
7	.22	1.0	.15	.13	.71	.67	.67	.67	.56	.33	.40	.44	$\overline{rp}, rt, \overline{re}_n$	\overline{lk}, re_b
8	.22	1.0	.15	.13	.67	.70	.89	.38	.30	.50	.40	.30	sh, rt, \overline{re}_n	$\overline{dp}, \overline{sc}$
9	.22	1.0	.15	.13	.71	.70	.78	.43	.40	.50	.10	.11	$\overline{sh}, rt, \overline{re}_n$	\overline{dp}, re_b
10	.22	1.0	.15	.13	.71	.70	.75	.57	.30	.50	.50	.25	lk, rt, \overline{re}_n	ge_a, \overline{sc}
Averages														
	.24	1.0	.16	.13	.72	.69	.81	.53	.57	.39	.28	.20		

APPENDIX B**EVALUATION RESULTS: ADDITIONAL TABLES FOR
AVERAGE OVERLAP AND AVERAGE P@K**

TABLE 51: (Chapter 9.3.1, Coronavirus-Latest, Variant of Table 39 for Twitter-Latest seeds): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 45 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (^r) superscript (e.g., Google Random - G^r).

k	Average Overlap				Average P@K									
	GM	EM	G^r	M^r	E^r	M^r	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$
10	1.0	1.0	.18	.18	.64	.81	.69	.63	.63	.89	.65	.71		
20	1.0	1.0	.16	.15	.58	.73	.67	.66	.76	.73	.58	.58		
30	.60	.83	.14	.13	.70	.77	.66	.73	.71	.81	.58	.58		
40	.43	.89	.15	.13	.65	.76	.67	.64	.73	.81	.59	.59		
50	.42	.76	.14	.13	.72	.74	.66	.68	.73	.78	.56	.55		
60	.48	.73	.15	.12	.70	.76	.65	.66	.71	.76	.56	.55		
70	.42	.73	.14	.12	.69	.76	.64	.68	.72	.75	.53	.55		
80	.36	.68	.15	.12	.67	.78	.68	.66	.70	.76	.54	.57		
90	.32	.65	.15	.13	.70	.79	.68	.66	.68	.73	.55	.59		
100	.29	.65	.15	.13	.70	.79	.69	.64	.65	.72	.55	.55		
150	.21	.63	.16	.12	.69	.79	.68	.68	.72	.73	.55	.55		
200	-	.48	-	.12	-	.80	-	.65	-	.72	-	.55		
300	-	.31	-	.11	-	.77	-	.67	-	.68	-	.55		
All	.12	.08	.12	.08	.68	.65	.68	.65	.55	.55	.55	.55		

TABLE 52: (Chapter 9.3.1, 2018 World Cup, Supplements Table 46 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 46 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, G^r - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

K	Average Overlap		Average P@K			
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}
10	.60	.17	.78	.71	.63	.45
20	.40	.13	.78	.70	.79	.50
30	.31	.12	.79	.61	.74	.45
40	.23	.11	.76	.65	.54	.43
50	.19	.11	.65	.65	.45	.44
60	.16	.11	.63	.63	.45	.49
70	.15	.10	.63	.63	.58	.46
80	.14	.10	.64	.62	.52	.45
90	.12	.09	.65	.62	.49	.44
100	.09	.08	.66	.63	.47	.45
All	.07	.07	.62	.62	.44	.44

TABLE 53: (Chapter 9.3.1, 2018 World Cup-Latest, Variant of Table 52 for Twitter-Latest seeds): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 46 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

k	Average Overlap		Average P@K			
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}
10	.46	.14	.66	.69	.00	.11
20	.17	.10	.75	.71	.08	.12
30	.13	.09	.72	.68	.08	.16
40	.10	.07	.71	.68	.06	.14
50	.09	.06	.65	.65	.09	.15
60	.07	.05	.62	.65	.10	.14
70	.05	.04	.63	.62	.14	.15
80	.04	.04	.66	.62	.16	.15
84	.04	.04	.67	.61	.15	.15
All	.03	.03	.62	.62	.15	.15

TABLE 54: (Chapter 9.3.1, Hurricane Harvey (collected 2020), Supplements Table 47 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 47 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (^r) superscript (e.g., Google Random - G^r).

K	Average Overlap		Average P@K			
	GM	$G^r M^r$	G	G^r	M_G	M_G^r
10	.50	.14	.77	.28	.13	.06
20	.37	.13	.53	.47	.10	.11
30	.40	.11	.51	.37	.11	.11
40	.34	.12	.50	.41	.09	.10
50	.29	.12	.46	.41	.10	.10
60	.27	.12	.45	.42	.07	.11
70	.25	.13	.44	.42	.09	.12
80	.22	.13	.43	.40	.08	.10
90	.22	.13	.46	.43	.09	.09
100	.21	.13	.47	.42	.08	.09
150	.19	.14	.42	.42	.13	.10
All	.12	.12	.41	.41	.10	.10

TABLE 55: (Chapter 9.3.1, Hurricane Harvey (collected 2017), Supplements Table 48 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 47 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (^r) superscript (e.g., Google Random - G^r).

K	Average Overlap				Average P@K									
	GM	EM	G^r	M^r	E^r	M^r	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$
10	.68	.25	.27	.15	.90	.13	.74	.13	.56	.00	.24	.06		
20	.56	.22	.24	.09	.74	.25	.72	.23	.41	.12	.22	.15		
30	.41	.11	.23	.06	.71	.24	.71	.24	.34	.25	.18	.15		
40	.41	-	.22	-	.73	-	.71	-	.24	-	.17	-		
50	.30	-	.20	-	.73	-	.72	-	.21	-	.15	-		
All	.14	.02	.14	.02	.72	.25	.72	.25	.15	.15	.15	.15		

TABLE 56: (Chapter 9.3.1, Flint Water Crisis, Supplements Table 49 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 48 (first column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

K	Average Overlap		Average P@K			
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}
10	.47	.17	.94	.79	.60	.53
20	.39	.15	.99	.83	.99	.52
30	.40	.16	.81	.82	.78	.50
40	.42	.15	.73	.84	.49	.48
50	.39	.16	.77	.83	.64	.49
60	.35	.16	.85	.80	.61	.48
70	.33	.17	.87	.80	.60	.45
80	.31	.17	.83	.83	.56	.49
90	.30	.18	.81	.82	.53	.50
100	.29	.18	.71	.83	.43	.50
150	.24	.21	.90	.82	.57	.50
All	.19	.19	.82	.82	.48	.48

TABLE 57: (Chapter 9.3.1, 2014 Ebola Virus Outbreak, Supplements Table 50 by showing additional average P@K instead of only P@10): Average overlap (of top 10 of 2,049 QP combinations) and average P@K (of top 10 of 2,049 QP combinations) for K top seeds scored by 1 – 3 QP combinations. “-” represents cases when seed count < K. Fig. 48 (second column) visualizes this Table. **Key:** For **Average Overlap**, GM - average overlap between Google and Micro-collection, EM - expert and Micro-collection. For **Average P@K**, G - P@K of Google seeds with respect to Micro-collections, E - average P@K of expert with respect to Micro-collection. M_G/M_E : P@K of Micro-collection seeds with respect to Google/Expert seeds. Randomly selected (not selected using QP combination-assigned scores) seeds have the (r) superscript (e.g., Google Random - G^r).

K	Average Overlap				Average P@K									
	GM	EM	G^r	M^r	E^r	M^r	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$
10	.24	1.0	.16	.13	.72	.69	.81	.53	.57	.39	.28	.20		
20	.16	.50	.12	.07	.89	.57	.79	.58	.38	.24	.25	.23		
30	.13	.33	.09	.04	.88	.57	.77	.61	.25	.24	.22	.25		
40	.11	.09	.08	.03	.77	.56	.76	.60	.22	.23	.25	.25		
48	.08	-	.07	-	.80	-	.76	-	.24	-	.24	-		
All	.04	.02	.04	.02	.78	.60	.78	.60	.24	.24	.24	.24		

APPENDIX C**EVALUATION RESULTS: ADDITIONAL TABLES FOR
AVERAGE P@K FOR ADDITIONAL OVERLAP INTERVALS
OF QP COMBINATIONS**

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]		(.80, .90]	
	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M
10	.53	.62	.60	.63	.73	.70	.72	.68	.78	.73	.75	.69	.62	.32	.80	.92	.99	.59	-	-
20	.48	.51	.63	.65	.75	.71	.76	.76	.75	.81	.75	.74	.74	.63	-	-	.79	.75	-	-
30	.51	.52	.59	.59	.72	.67	.74	.69	.77	.84	.77	.77	.74	.64	.74	.62	.72	.55	.77	.81
40	.53	.45	.58	.57	.74	.69	.76	.69	.79	.81	.80	.77	.77	.69	.76	.66	.76	.67	.74	.61
50	.54	.45	.58	.55	.75	.69	.76	.64	.82	.76	.80	.79	.76	.68	.74	.69	.75	.67	.71	.67
60	.58	.49	.55	.51	.77	.70	.79	.68	.80	.73	.76	.72	.74	.70	.72	.69	-	-	-	-
70	.54	.46	.55	.50	.78	.72	.77	.66	.78	.71	.78	.75	.77	.71	.77	.70	-	-	-	-
80	.56	.49	.55	.49	.78	.72	.78	.64	.79	.70	.78	.72	.78	.72	.76	.73	-	-	-	-
90	.58	.51	.53	.47	.78	.71	.79	.65	.80	.69	.79	.73	.79	.73	.78	.66	-	-	-	-
100	.57	.51	.54	.48	.77	.70	.78	.65	.81	.69	.78	.71	.78	.69	.77	.67	-	-	-	-
150	.55	.53	.57	.49	.77	.67	.78	.66	.81	.66	.80	.69	.80	.70	.84	.78	-	-	-	-
200	.57	.54	.60	.52	.76	.64	.77	.68	.78	.68	.80	.68	.83	.65	-	-	-	-	-	-
300	-	-	.64	.54	.73	.63	.76	.66	.76	.60	-	-	-	-	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]		(.80, .90]	
	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M	E	M
10	898	906	416	416	201	212	65	65	90	90	165	165	10	10	93	93	10	10	-	-
20	611	611	616	630	318	311	151	148	119	110	117	116	79	79	-	-	4	4	-	-
30	362	362	877	867	361	344	151	147	108	105	89	83	38	38	24	24	24	24	5	5
40	220	220	958	933	414	404	193	192	75	71	64	64	44	44	56	56	24	24	1	1
50	185	185	939	907	476	472	155	151	73	73	127	127	65	65	20	20	6	6	1	1
60	148	148	909	877	563	555	114	114	160	160	91	91	54	54	10	10	-	-	-	-
70	127	127	908	887	569	564	152	152	139	139	102	102	31	31	21	21	-	-	-	-
80	118	118	906	899	584	582	152	152	163	163	82	82	42	42	2	2	-	-	-	-
90	108	108	879	879	627	627	155	155	156	156	103	103	17	17	4	4	-	-	-	-
100	104	104	905	905	615	615	162	162	136	136	102	102	23	23	2	2	-	-	-	-
150	73	73	1,007	1,007	523	523	182	182	160	160	85	85	18	18	1	1	-	-	-	-
200	23	23	1,201	1,201	345	345	262	262	192	192	24	24	2	2	-	-	-	-	-	-
300	-	-	1,394	1,394	578	578	74	74	3	3	-	-	-	-	-	-	-	-	-	-

TABLE 58: (Chapter 9.3.1 & 9.3.2, Coronavirus-Expert reference (E), Supplements Table 37 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	.55	.32	.70	.47	.71	.56	.72	.48	.84	.46	.84	.68	.78	.69	-	-	.67	.33
20	.57	.38	.62	.37	.66	.56	.76	.69	.76	.75	1.00	.83	-	-	-	-	-	-
30	.58	.36	.60	.41	.68	.58	.74	.71	.78	.75	-	-	-	-	-	-	-	-
40	.58	.31	.63	.41	.69	.57	.72	.58	-	-	-	-	-	-	-	-	-	-
50	.62	.30	.65	.40	.67	.54	.62	.27	-	-	-	-	-	-	-	-	-	-
60	.62	.29	.66	.41	.66	.53	-	-	-	-	-	-	-	-	-	-	-	-
70	.65	.32	.67	.42	.65	.52	-	-	-	-	-	-	-	-	-	-	-	-
80	.62	.36	.67	.43	.65	.55	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	.66	.44	.65	.50	-	-	-	-	-	-	-	-	-	-	-	-
100	-	-	.65	.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-
200	-	-	.62	.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	1,004	990	426	422	377	349	171	159	40	37	24	19	5	5	-	-	2	2
20	675	675	576	576	511	511	233	233	51	51	3	3	-	-	-	-	-	-
30	450	450	911	911	464	464	215	215	9	9	-	-	-	-	-	-	-	-
40	221	221	1,253	1,253	538	538	37	37	-	-	-	-	-	-	-	-	-	-
50	80	80	1,498	1,498	470	470	1	1	-	-	-	-	-	-	-	-	-	-
60	43	43	1,666	1,666	340	340	-	-	-	-	-	-	-	-	-	-	-	-
70	23	23	1,766	1,766	260	260	-	-	-	-	-	-	-	-	-	-	-	-
80	14	14	1,910	1,910	125	125	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	1,985	1,985	64	64	-	-	-	-	-	-	-	-	-	-	-	-
100	-	-	2,049	2,049	-	-	-	-	-	-	-	-	-	-	-	-	-	-
200	-	-	2,049	2,049	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE 59: (Chapter 9.3.1 & 9.3.2, 2018 World Cup, Supplements Table 46 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	G	M	G	M	G	M	G	M	G	M	G	M
10	.36	.13	.47	.12	.57	.11	.62	.11	.67	.18	.74	.14
20	.40	.11	.40	.11	.50	.11	.53	.09	.55	.12	-	-
30	.38	.09	.40	.11	.50	.11	.51	.10	.51	.09	-	-
40	.38	.08	.40	.11	.50	.11	.50	.09	.48	.09	-	-
50	.40	.08	.39	.10	.49	.11	.48	.09	-	-	-	-
60	.42	.08	.40	.10	.49	.11	.47	.09	-	-	-	-
70	.43	.08	.41	.09	.47	.11	.48	.08	-	-	-	-
80	.42	.08	.41	.09	.47	.11	.45	.09	-	-	-	-
90	.43	.08	.41	.09	.46	.11	.48	.09	-	-	-	-
100	.43	.09	.41	.09	.45	.10	.46	.10	-	-	-	-
150	.44	.09	.41	.09	.43	.10	-	-	-	-	-	-
300	-	-	-	-	.41	.10	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	G	M	G	M	G	M	G	M	G	M	G	M
10	731	730	618	573	451	446	169	169	62	62	18	18
20	425	424	734	684	447	447	340	340	103	103	-	-
30	259	258	805	771	557	557	340	340	88	88	-	-
40	152	152	977	977	530	530	347	347	43	43	-	-
50	107	107	1,023	1,023	594	594	325	325	-	-	-	-
60	81	81	1,111	1,111	646	646	211	211	-	-	-	-
70	57	57	1,165	1,165	740	740	87	87	-	-	-	-
80	27	27	1,182	1,182	806	806	34	34	-	-	-	-
90	22	22	1,192	1,192	782	782	53	53	-	-	-	-
100	17	17	1,089	1,089	879	879	64	64	-	-	-	-
150	10	10	735	735	1,304	1,304	-	-	-	-	-	-
300	-	-	-	-	2,049	2,049	-	-	-	-	-	-

TABLE 60: (Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2020), Supplements Table 47 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]		(.80, .90]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	.56	.16	.65	.20	.70	.26	.70	.25	.74	.34	.77	.47	.75	.31	.85	.57	.94	.60	-	-
20	.65	.14	.67	.14	.69	.21	.73	.31	.70	.41	.72	.43	.68	.39	.80	.50	-	-	.79	.50
30	.65	.07	.70	.13	.73	.21	.72	.28	.70	.33	.72	.33	-	-	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]		(.50, .60]		(.60, .70]		(.70, .80]		(.80, .90]	
	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M
10	423	422	450	450	481	481	323	323	230	230	98	98	35	35	3	3	6	6	-	-
20	210	210	510	510	624	624	371	371	272	272	57	57	3	3	1	1	-	-	1	1
30	94	94	601	601	599	599	557	557	191	191	7	7	-	-	-	-	-	-	-	-

TABLE 61: (Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2017), Supplements Table 48 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]	
	E	M	E	M	E	M	E	M	E	M		E	M	E	M	E	M	E	M		
10	.29	.24	.19	.26	.15	.32	.08	.00	.30	.00	10	1,512	1,511	443	443	84	84	8	8	2	2
20	.29	.24	.22	.23	.16	.12	.27	.15	-	-	20	1,030	1,030	991	991	23	23	5	5	-	-
30	.27	.22	.26	.20	.20	.26	-	-	-	-	30	441	441	1,602	1,602	6	6	-	-	-	-
37	.25	.23	.25	.19	-	-	-	-	-	-	37	301	301	1,748	1,748	-	-	-	-	-	-

TABLE 62: (Chapter 9.3.1 & 9.3.2, Hurricane Harvey (collected 2017)-Expert reference (E), Supplements Table 48 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (left) of QP combinations of different overlap ranges and the count of QP combinations (right) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	G	M	G	M	G	M	G	M	G	M	G	M
10	.60	.44	.75	.58	.81	.56	.92	.68	.91	.55	.95	.64
20	.71	.52	.65	.45	.85	.66	.85	.55	.90	.63	1.0	1.0
30	.73	.49	.66	.43	.86	.66	.83	.53	.81	.52	.86	.91
40	.75	.48	.73	.43	.85	.65	.81	.53	.81	.53	.74	.57
50	.80	.47	.76	.40	.86	.65	.82	.53	.78	.53	-	-
60	.81	.46	.78	.39	.86	.63	.81	.53	.76	.55	-	-
70	.80	.44	.79	.37	.87	.62	.81	.53	.85	.63	-	-
80	.81	.42	.79	.38	.87	.59	.81	.53	.83	.56	-	-
90	.81	.41	.81	.38	.86	.56	.82	.52	.76	.47	-	-
100	.82	.42	.81	.37	.86	.55	.82	.53	-	-	-	-
150	-	-	.83	.45	.83	.48	.84	.56	-	-	-	-
500	-	-	-	-	.82	.48	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	G	M	G	M	G	M	G	M	G	M	G	M
10	597	588	589	584	505	499	282	282	65	65	11	11
20	299	290	743	732	643	643	280	280	83	83	1	1
30	207	205	737	734	681	681	288	288	131	131	5	5
40	132	132	773	773	715	715	309	309	98	98	22	22
50	82	82	767	767	749	749	359	359	92	92	-	-
60	47	47	772	772	793	793	366	366	71	71	-	-
70	27	27	753	753	808	808	424	424	37	37	-	-
80	18	18	708	708	891	891	423	423	9	9	-	-
90	15	15	623	623	1,034	1,034	375	375	2	2	-	-
100	10	10	523	523	1,153	1,153	363	363	-	-	-	-
150	-	-	123	123	1,668	1,668	258	258	-	-	-	-
500	-	-	-	-	2,049	2,049	-	-	-	-	-	-

TABLE 63: (Chapter 9.3.1 & 9.3.2, Flint water crisis, Supplements Table 49 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (Top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]	
	G	M	G	M	G	M	G	M
10	.75	.35	.78	.42	.77	.34	.72	.59
20	.74	.27	.80	.27	.81	.35	-	-
30	.77	.23	.80	.24	.88	.27	-	-
40	.80	.24	.82	.23	.80	.22	-	-
48	.83	.24	.81	.24	-	-	-	-
100	-	-	.78	.24	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]	
	G	M	G	M	G	M	G	M
10	1,155	1,155	475	475	408	408	11	11
20	774	774	1,180	1,180	95	95	-	-
30	598	598	1,437	1,437	14	14	-	-
40	370	370	1,619	1,619	60	60	-	-
48	188	188	1,861	1,861	-	-	-	-
100	-	-	2,049	2,049	-	-	-	-

TABLE 64: (Chapter 9.3.1 & 9.3.2, 2014 Ebola Virus Outbreak, Supplements Table 50 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (left) of QP combinations of different overlap ranges and the count of QP combinations (right) that produced the average.

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	E	M	E	M	E	M	E	M	E	M	E	M
10	.62	.35	.64	.57	.67	.42	.73	.37	.61	.32	.58	.29
20	.64	.28	.61	.30	.64	.22	.56	.22	.52	.22	.54	.24
30	.69	.26	.57	.22	.53	.22	.53	.24	.58	.24	-	-
40	.70	.23	.57	.23	-	-	-	-	-	-	-	-

k	0		(.00, .10]		(.10, .20]		(.20, .30]		(.30, .40]		(.40, .50]	
	E	M	E	M	E	M	E	M	E	M	E	M
10	1,310	1,310	118	118	183	183	95	95	138	138	98	98
20	1,013	1,013	511	511	241	241	58	58	147	147	79	79
30	809	809	919	919	214	214	93	93	14	14	-	-
40	717	717	1,332	1,332	-	-	-	-	-	-	-	-

TABLE 65: (Chapter 9.3.1 & 9.3.2, 2014 Ebola virus outbreak-Expert reference (E), Supplements Table 50 by providing average P@K for additional overlap ranges of QP combinations; not just the top 10 overlap): Average P@K (top) of QP combinations of different overlap ranges and the count of QP combinations (bottom) that produced the average.

APPENDIX D

EVALUATION RESULTS: SUPPLEMENTARY LINE VIS FOR OVERLAP AND P@K FOR DIFFERENT COMBINATIONS OF QUALITY PROXIES

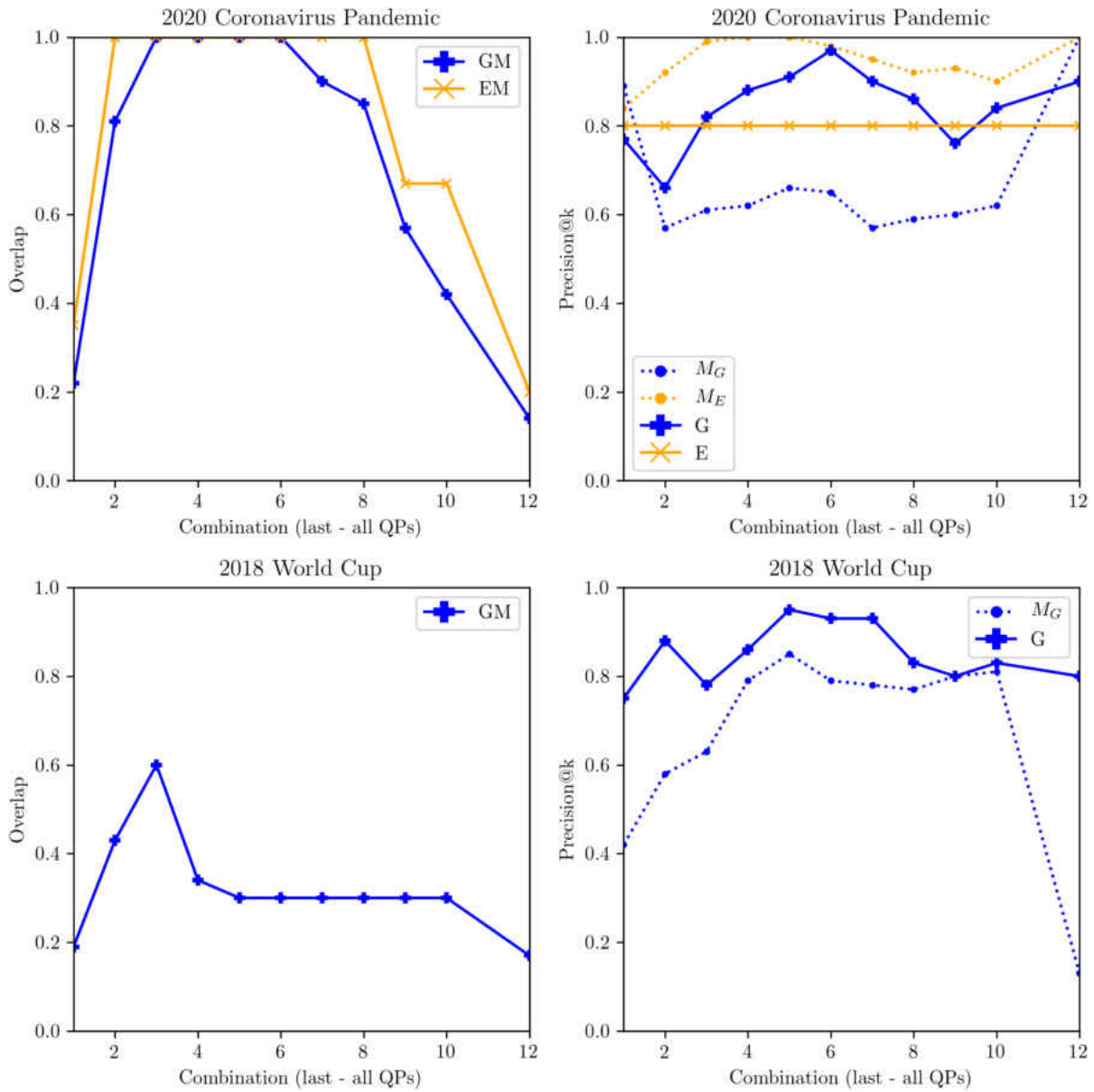


Fig. 52: (Chapter 9.3.1, Coronavirus and 2018 World Cup, Supplementary Line chart visualization of Table 42 (Coronavirus) and Table 66): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies

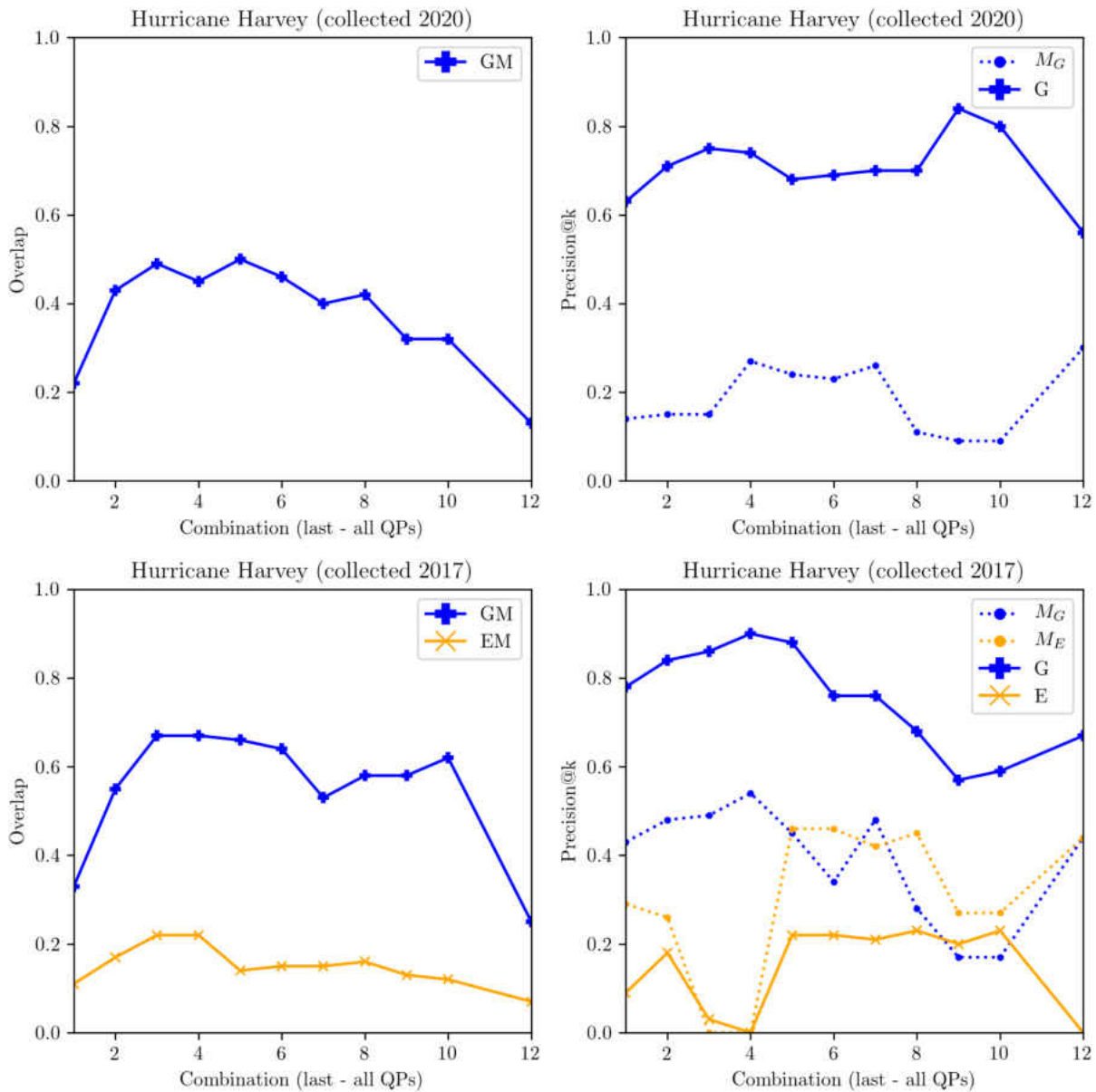


Fig. 53: (Chapter 9.3.1, Hurricane Harvey (collected 2020 and 2017), Supplementary Line chart visualization of Table 67 (collected 2020) and Table 68 (collected 2017)): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies

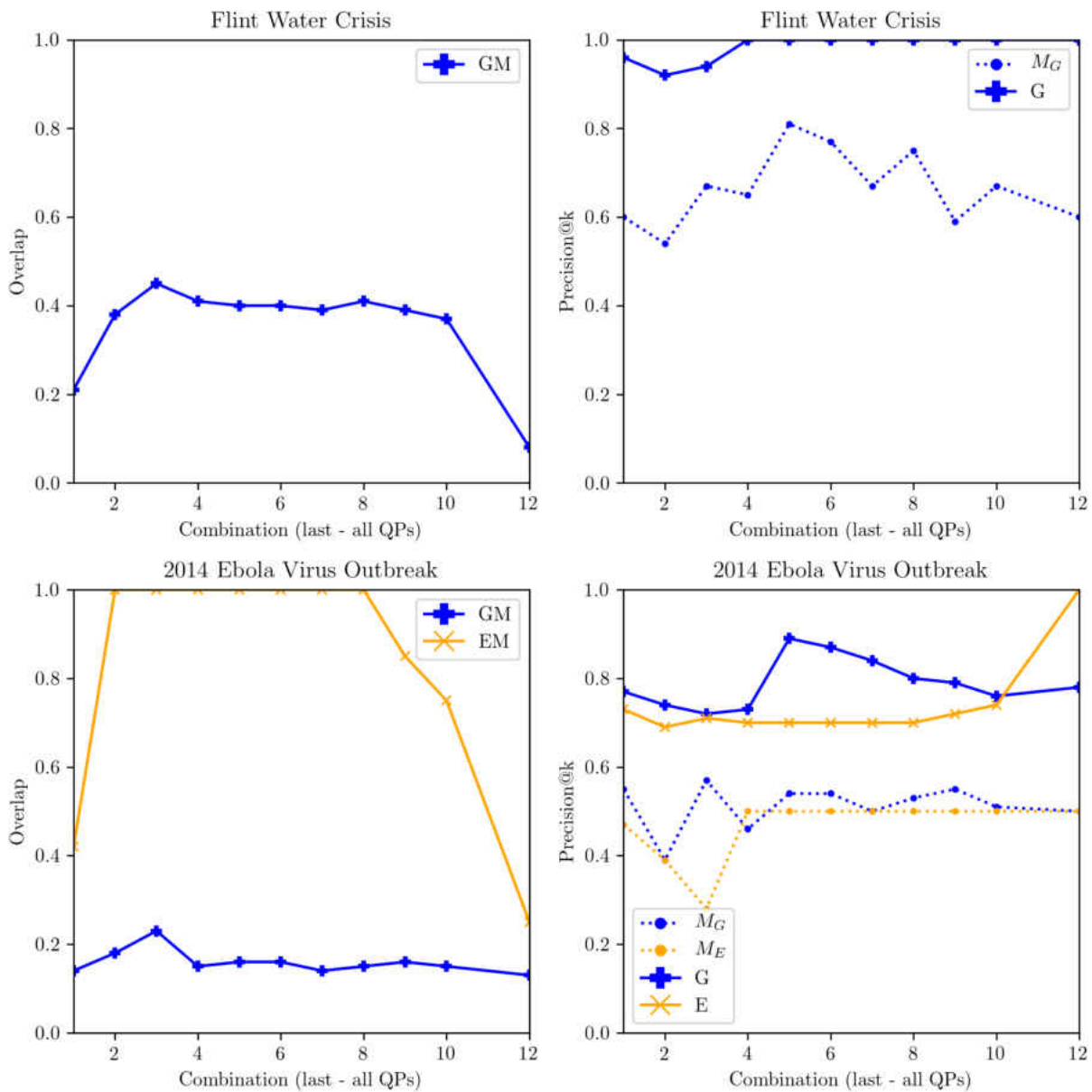


Fig. 54: (Chapter 9.3.1, Flint Water Crisis and 2014 Ebola Virus Outbreak, Supplementary Line chart visualization of Table 69 (Flint Water Crisis) and Table 70 (2014 Ebola Virus Outbreak)): Overlap (first column) and Precision@K (second column) for different QP Combinations. The last x-value is the selection of all Quality Proxies

APPENDIX E

**EVALUATION RESULTS: ADDITIONAL TABLES FOR
AVERAGE OVERLAP AND AVERAGE P@10 FOR
DIFFERENT COMBINATIONS OF QUALITY PROXIES**

TABLE 66: 4. (Chapter 9.3.1, 2018 World Cup, Supplement Table 46 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combs	Average Overlap		Average P@10			
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}
1	.19	.06	.75	.43	.42	.45
2	.43	.13	.88	.62	.58	.60
3	.60	.16	.78	.77	.63	.51
4	.34	.17	.86	.63	.79	.48
5	.30	.16	.95	.65	.85	.54
6	.30	.17	.93	.74	.79	.47
7	.30	.19	.93	.59	.78	.49
8	.30	.17	.83	.69	.77	.52
9	.30	.18	.80	.69	.80	.59
10	.30	.19	.83	.59	.81	.41
All	.17	.00	.80	.75	.13	.13

TABLE 67: 4. (Chapter 9.3.1, Hurricane Harvey (collected 2020), Supplement Table 47 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combs	Average Overlap		Average P@10			
	GM	$G^r M^r$	G	G^r	M_G	M_{G^r}
1	.22	.06	.63	.35	.14	.14
2	.43	.09	.71	.46	.15	.14
3	.49	.14	.75	.33	.15	.12
4	.45	.15	.74	.45	.27	.05
5	.50	.13	.68	.35	.24	.12
6	.46	.15	.69	.42	.23	.10
7	.40	.13	.70	.44	.26	.09
8	.42	.14	.70	.42	.11	.06
9	.32	.15	.84	.43	.09	.09
10	.32	.13	.80	.39	.09	.08
All	.13	.00	.56	.56	.30	.00

TABLE 68: 4. (Chapter 9.3.1, Hurricane Harvey (collected 2017), Supplement Table 48 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combs	Average Overlap				Average P@10							
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	M_{G^r}	M_{E^r}
1	.33	.11	.10	.05	.78	.09	.70	.35	.43	.29	.16	.15
2	.55	.17	.18	.09	.84	.18	.73	.26	.48	.26	.23	.15
3	.67	.22	.28	.15	.86	.03	.73	.19	.49	.00	.26	.12
4	.67	.22	.29	.15	.90	.00	.76	.23	.54	.00	.28	.15
5	.66	.14	.27	.14	.88	.22	.71	.32	.45	.46	.18	.13
6	.64	.15	.29	.15	.76	.22	.72	.20	.34	.46	.18	.15
7	.53	.15	.27	.14	.76	.21	.71	.21	.48	.42	.27	.16
8	.58	.16	.29	.14	.68	.23	.68	.27	.28	.45	.23	.06
9	.58	.13	.26	.15	.57	.20	.78	.25	.17	.27	.28	.09
10	.62	.12	.28	.15	.59	.23	.71	.21	.17	.27	.24	.11
All	.25	.07	.06	.00	.67	.00	.70	.00	.44	.44	.17	.13

TABLE 69: 4. (Chapter 9.3.1, Flint water crisis, Supplement Table 49 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combs	Average Overlap		Average P@10			
	GM	$G^r M^r$	G	G^r	M_G	$M_{G^r}^r$
1	.21	.05	.96	.80	.60	.56
2	.38	.12	.92	.78	.54	.46
3	.45	.15	.94	.77	.67	.64
4	.41	.14	1.0	.82	.65	.59
5	.40	.15	1.0	.91	.81	.48
6	.40	.16	1.0	.85	.77	.51
7	.39	.18	1.0	.84	.67	.53
8	.41	.17	1.0	.84	.75	.59
9	.39	.16	1.0	.83	.59	.62
10	.37	.14	1.0	.85	.67	.50
All	.08	.00	1.0	.70	.60	.56

TABLE 70: 4. (Chapter 9.3.1, 2014 Ebola Virus Outbreak, Supplement Table 50 by varying combinations): Average overlap and average P@10 for 2,100 top scoring combinations of 1-, 2-, 3-,..., 9-, and 10-combinations of QPs, and additionally selecting all QPs (last row).

Combs	Average Overlap				Average P@10							
	GM	EM	$G^r M^r$	$E^r M^r$	G	E	G^r	E^r	M_G	M_E	$M_{G^r}^r$	$M_{E^r}^r$
1	.14	.42	.01	.09	.77	.73	.81	.56	.55	.47	.27	.23
2	.18	1.0	.12	.12	.74	.69	.83	.61	.39	.39	.23	.16
3	.23	1.0	.15	.13	.72	.71	.84	.53	.57	.28	.27	.18
4	.15	1.0	.15	.13	.73	.70	.81	.63	.46	.50	.29	.24
5	.16	1.0	.16	.13	.89	.70	.73	.61	.54	.50	.26	.29
6	.16	1.0	.17	.14	.87	.70	.72	.61	.54	.50	.23	.14
7	.14	1.0	.17	.13	.84	.70	.78	.59	.50	.50	.22	.22
8	.15	1.0	.16	.14	.80	.70	.80	.50	.53	.50	.16	.28
9	.16	.85	.16	.13	.79	.72	.76	.49	.55	.50	.28	.25
10	.15	.75	.17	.14	.76	.74	.77	.54	.51	.50	.15	.23
All	.13	.25	.00	.00	.78	1.0	.88	.56	.50	.50	.00	.30

APPENDIX F

**EVALUATION RESULTS: EMPIRICAL CUMULATIVE
DISTRIBUTION FUNCTION OF THE DIVERSITY (D_U -
UNIQUE RATIO) OF REFERENCE AND
MICRO-COLLECTION SEEDS**

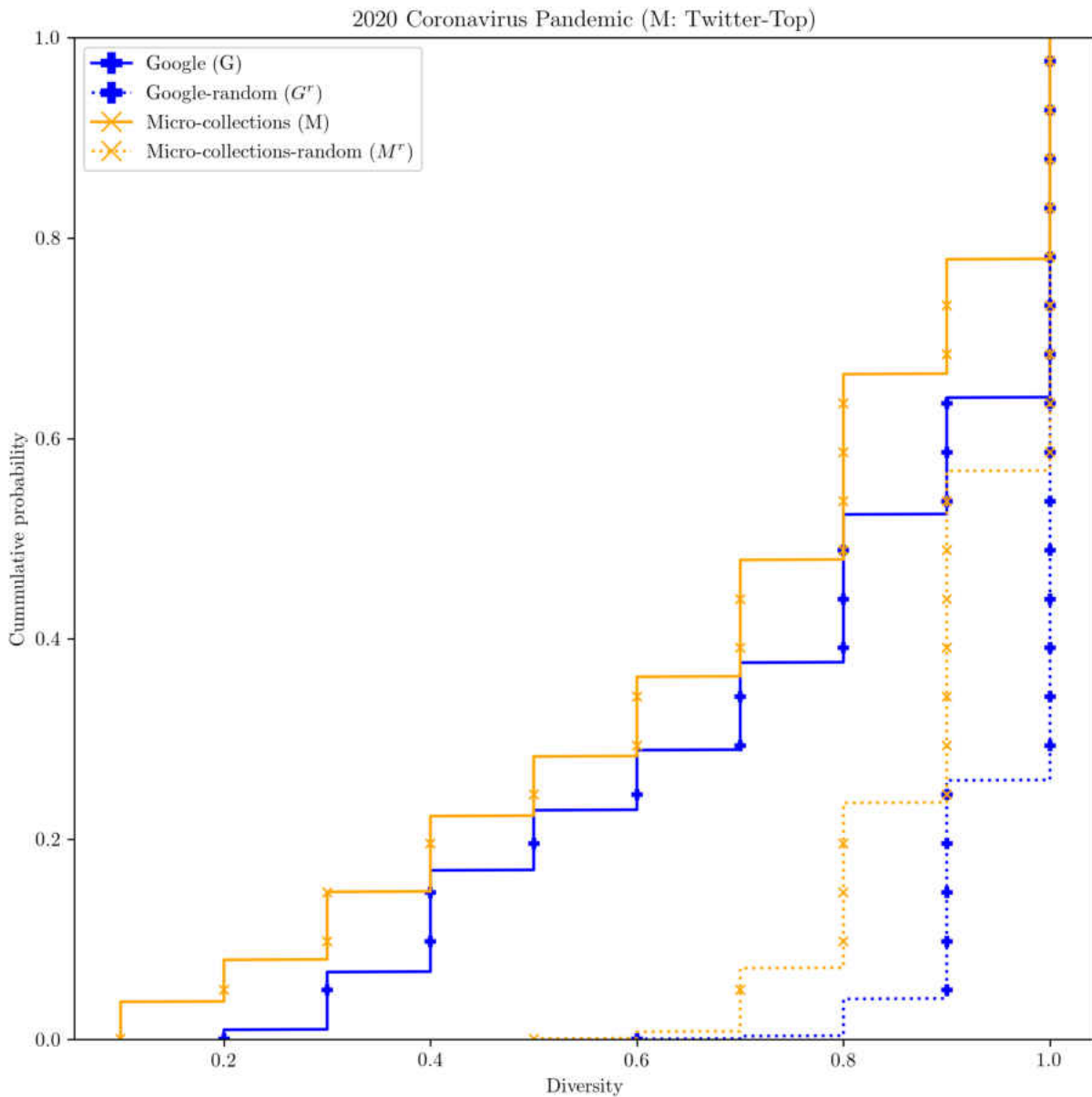


Fig. 55: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

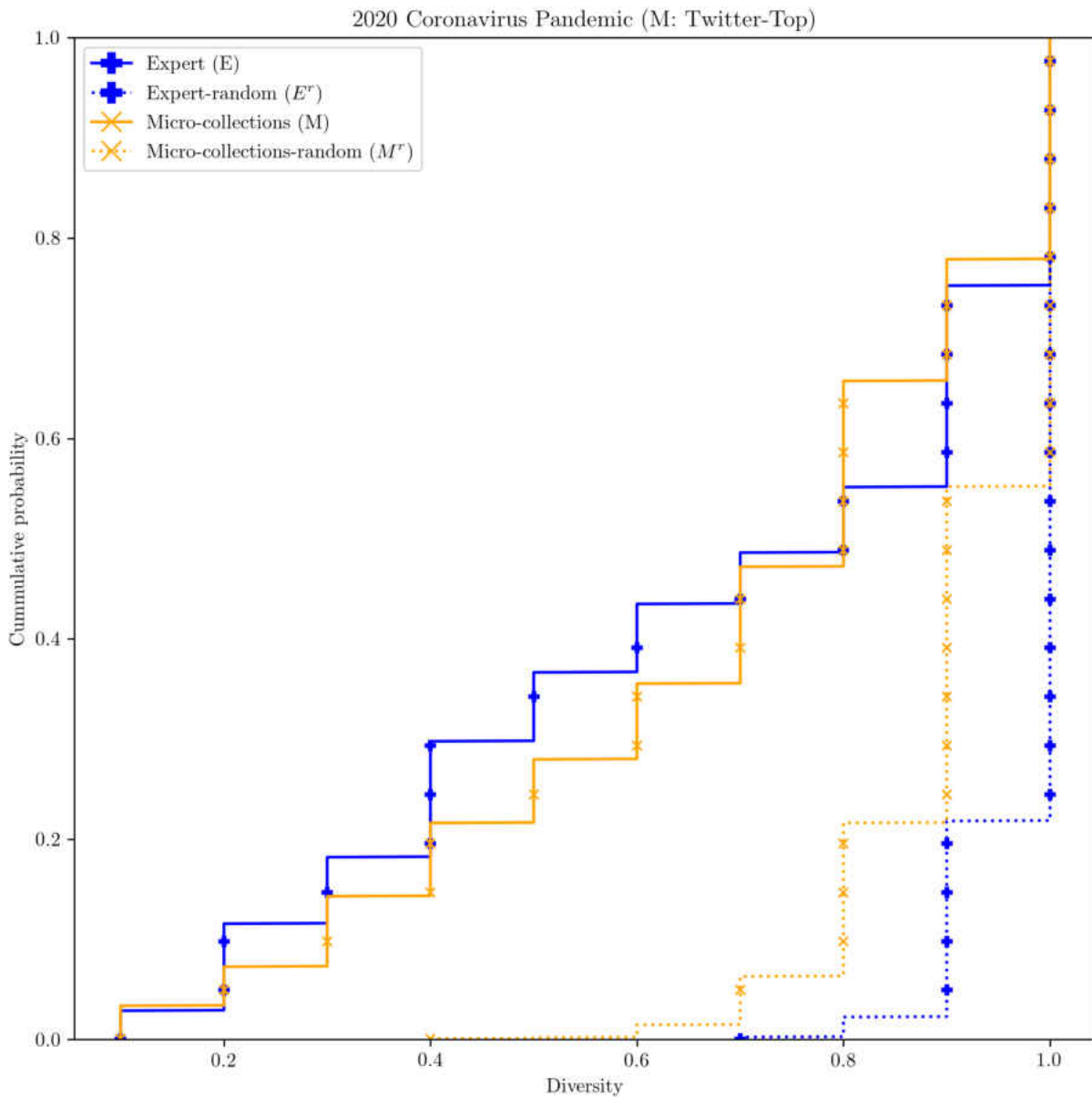


Fig. 56: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

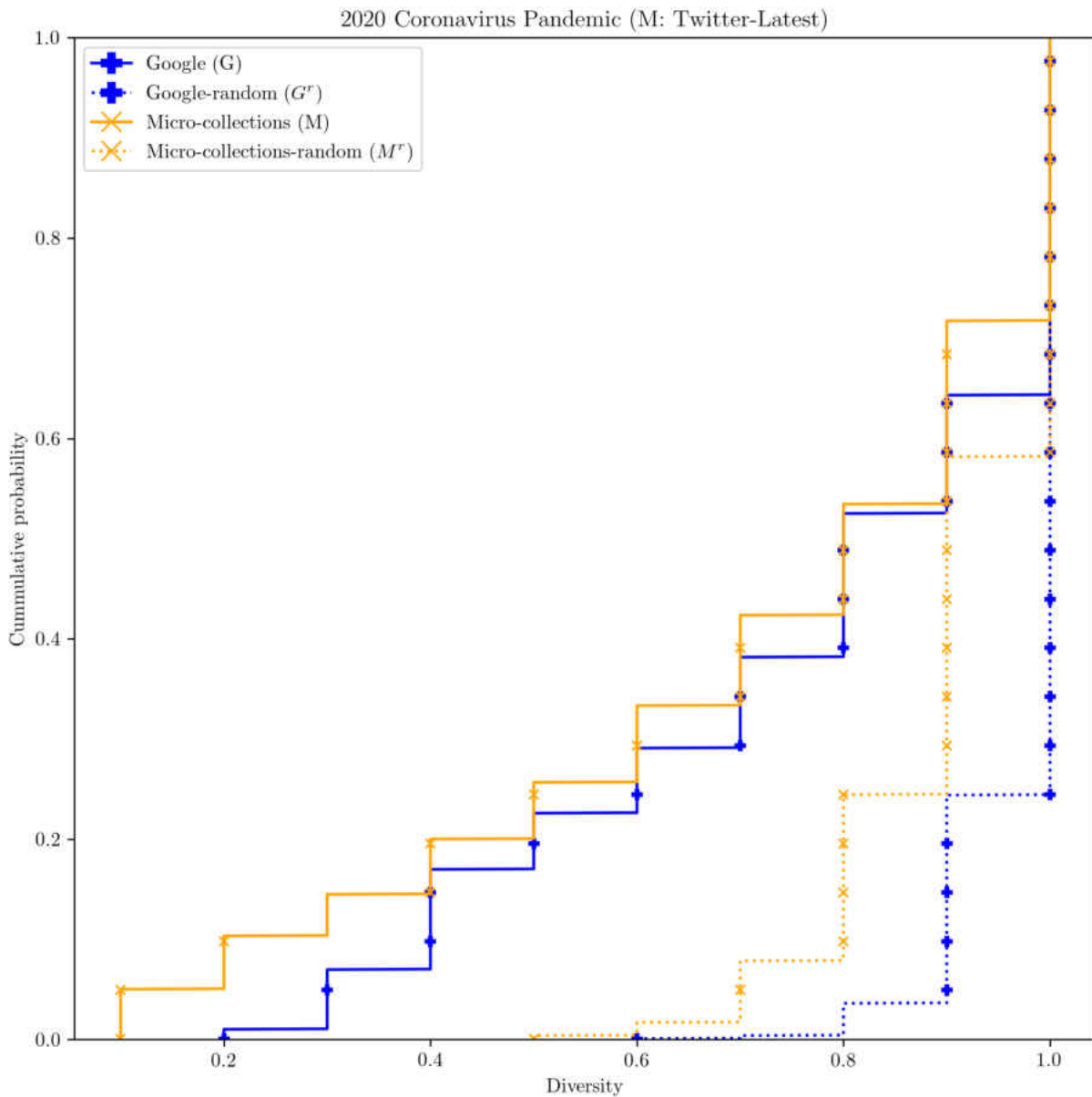


Fig. 57: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

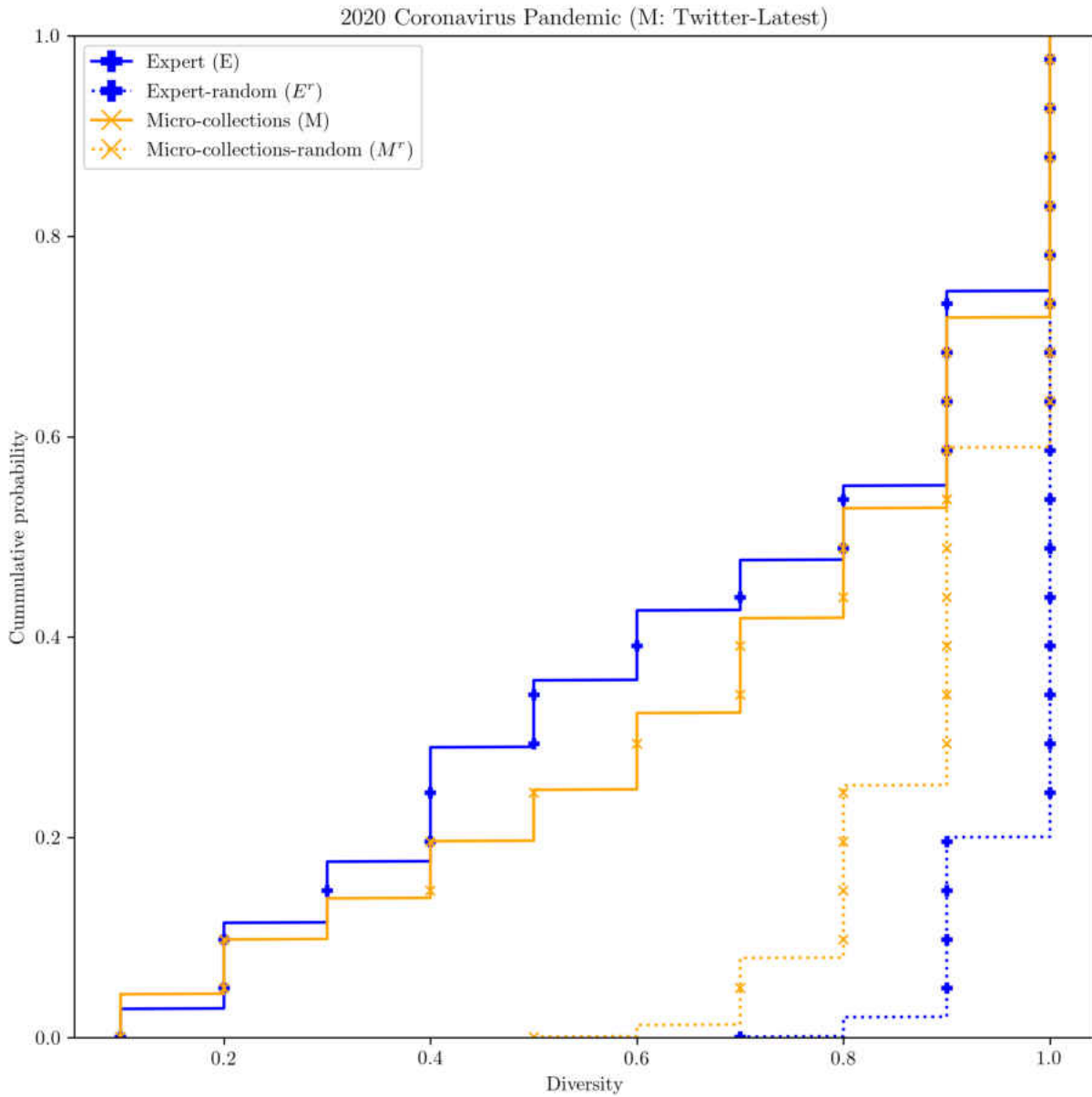


Fig. 58: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

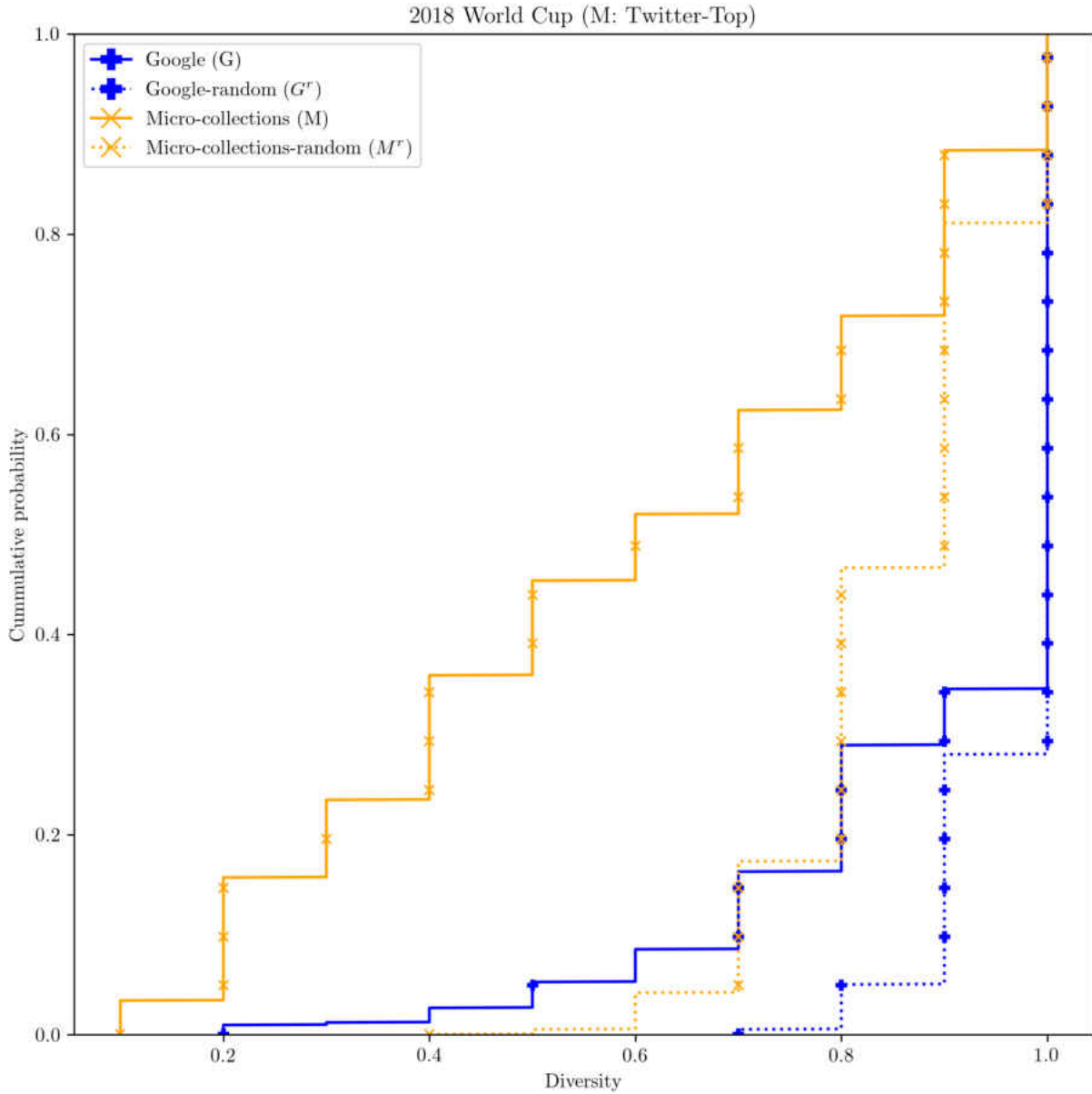


Fig. 59: (Chapter 9.3.4, 2018 World Cup-Top, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

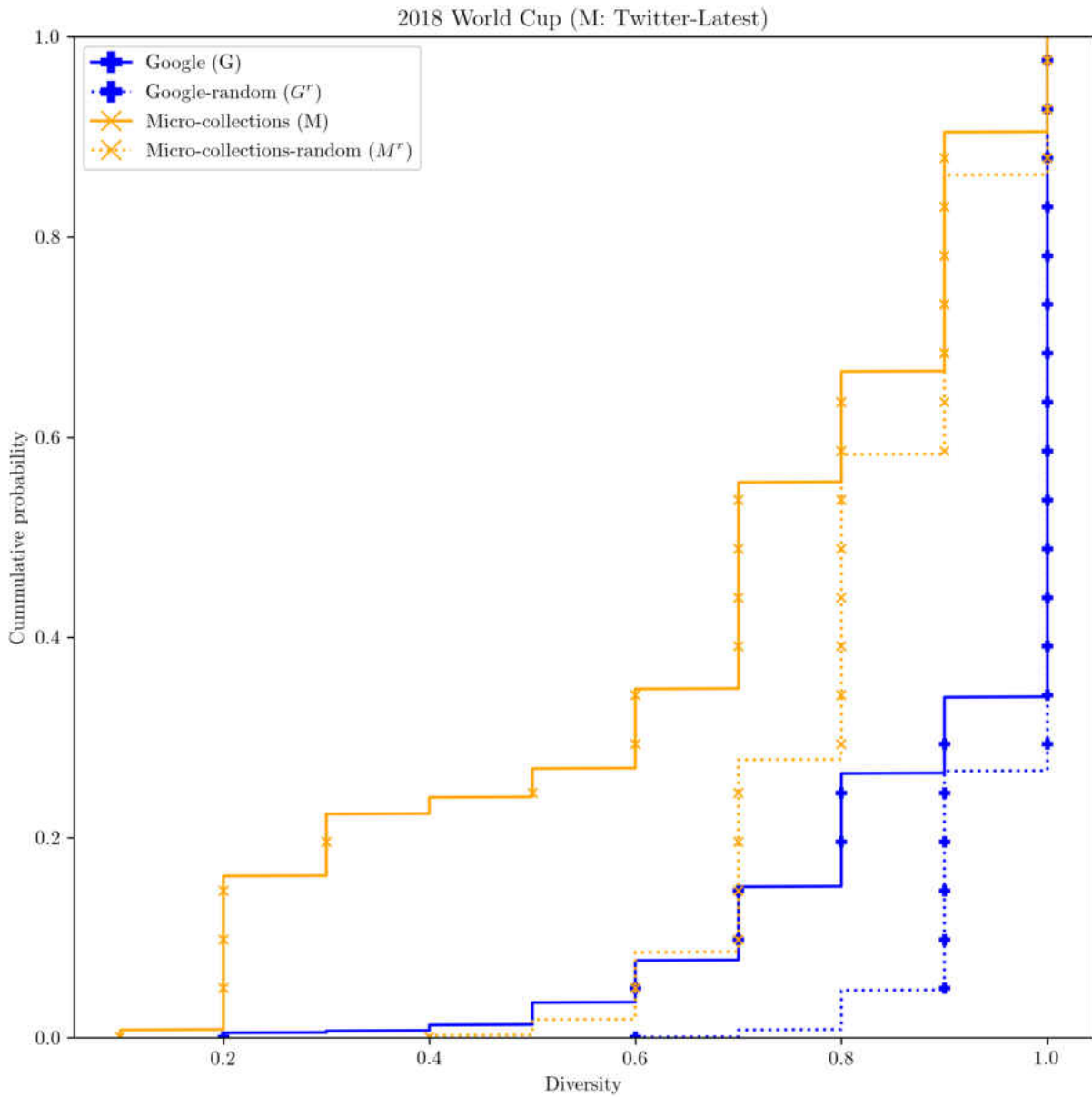


Fig. 60: (Chapter 9.3.4, 2018 World Cup-Latest, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

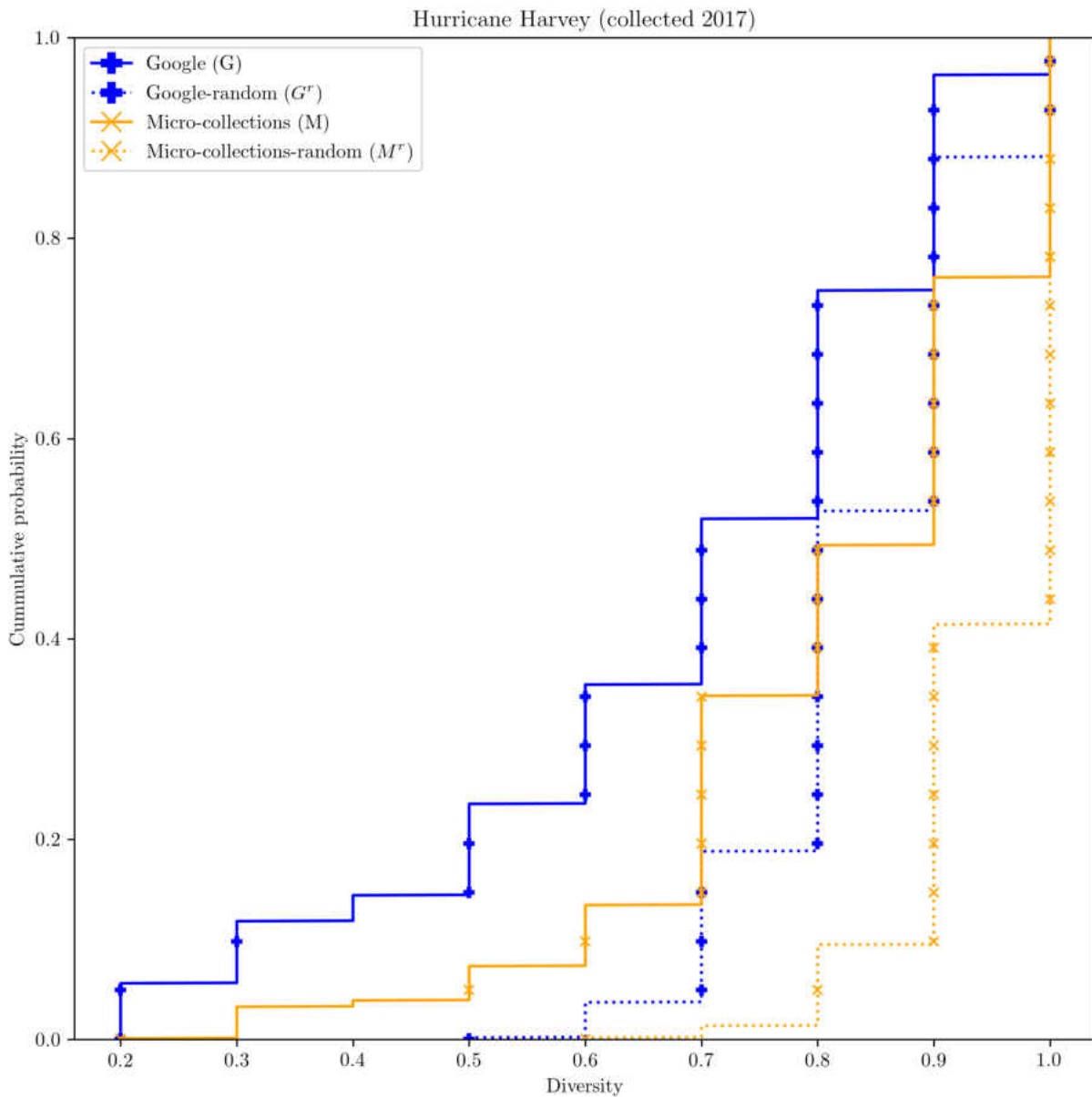


Fig. 61: (Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

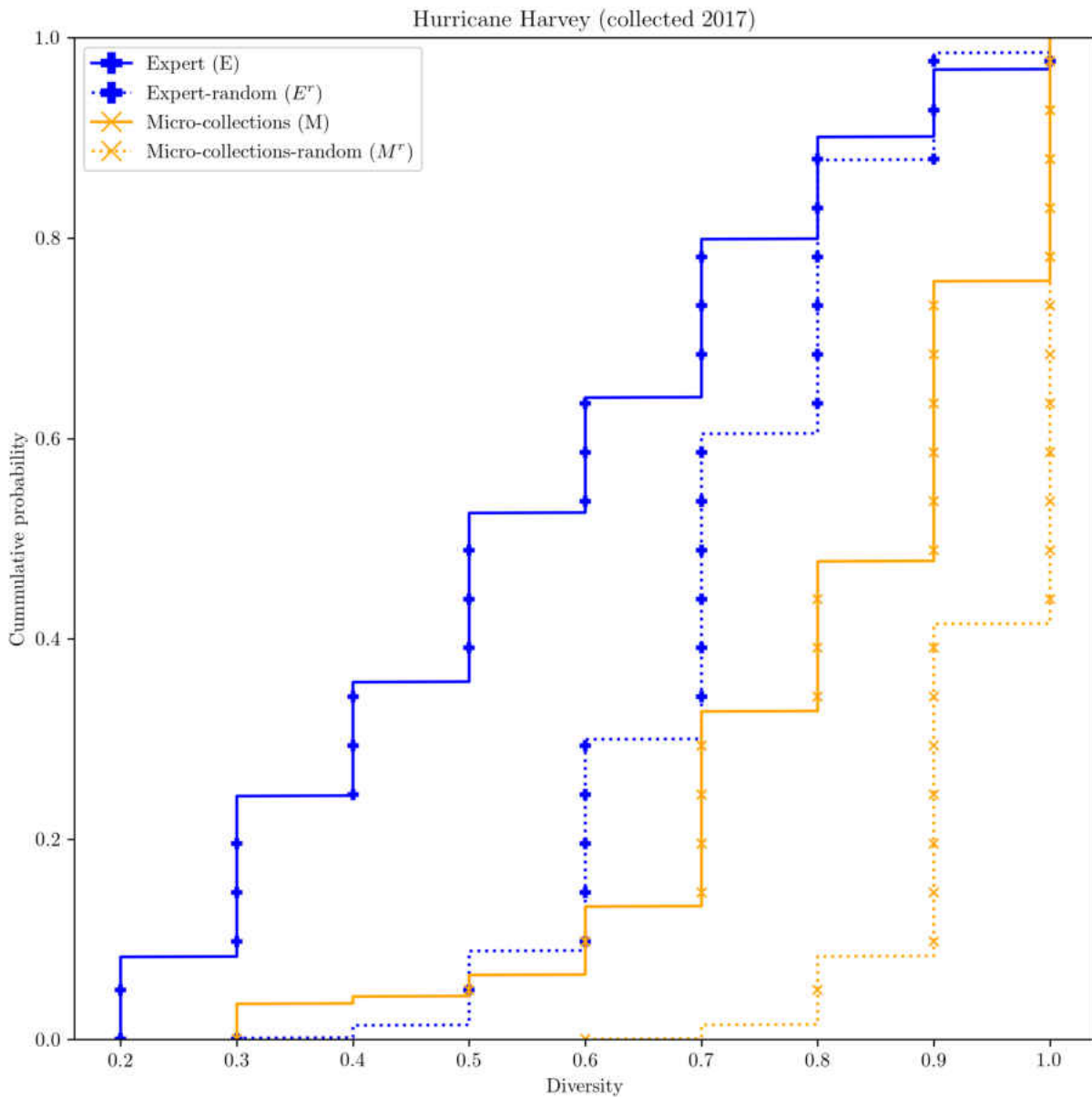


Fig. 62: (Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

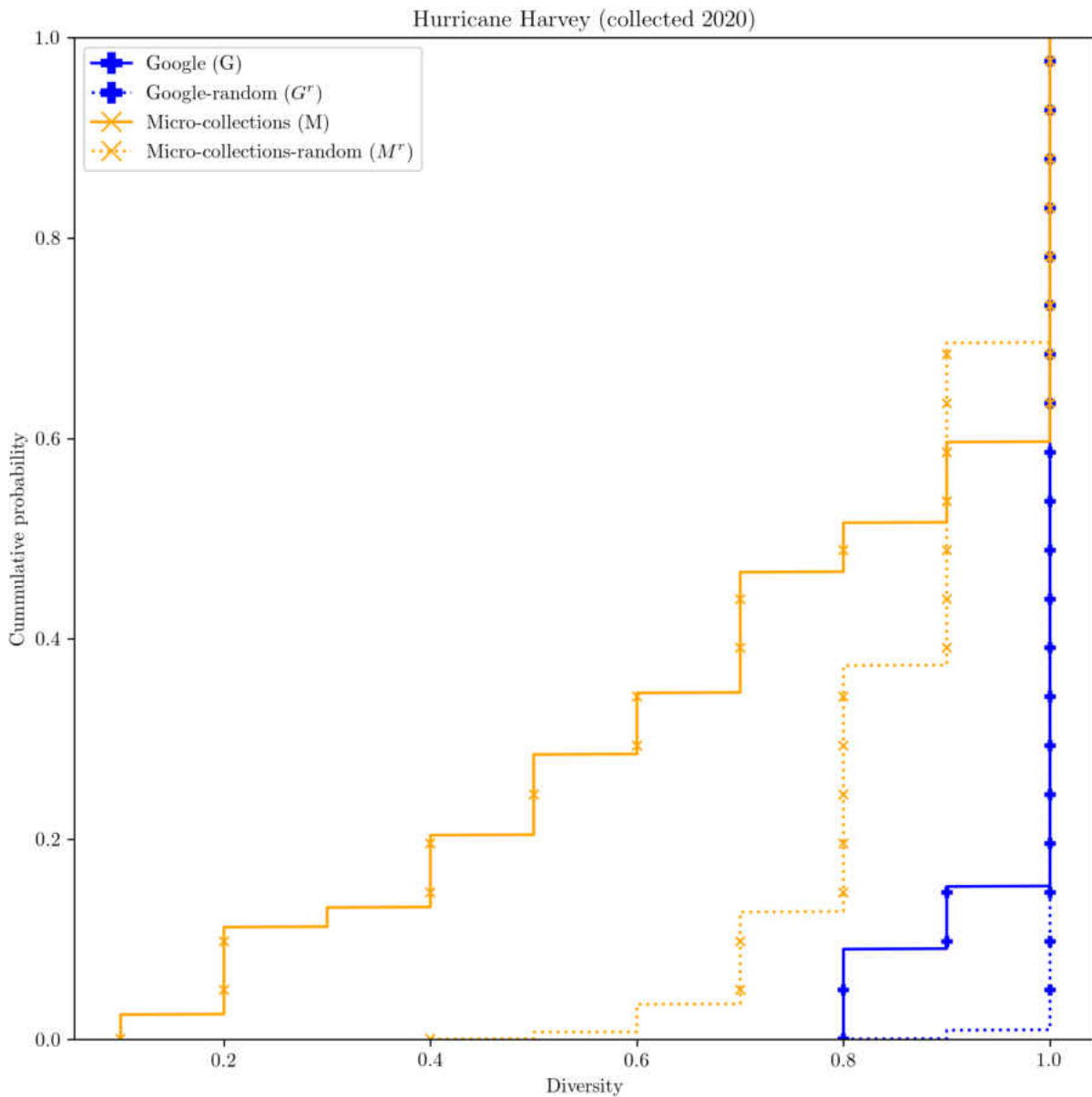


Fig. 63: (Chapter 9.3.4, Hurricane Harvey (collected 2020), Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

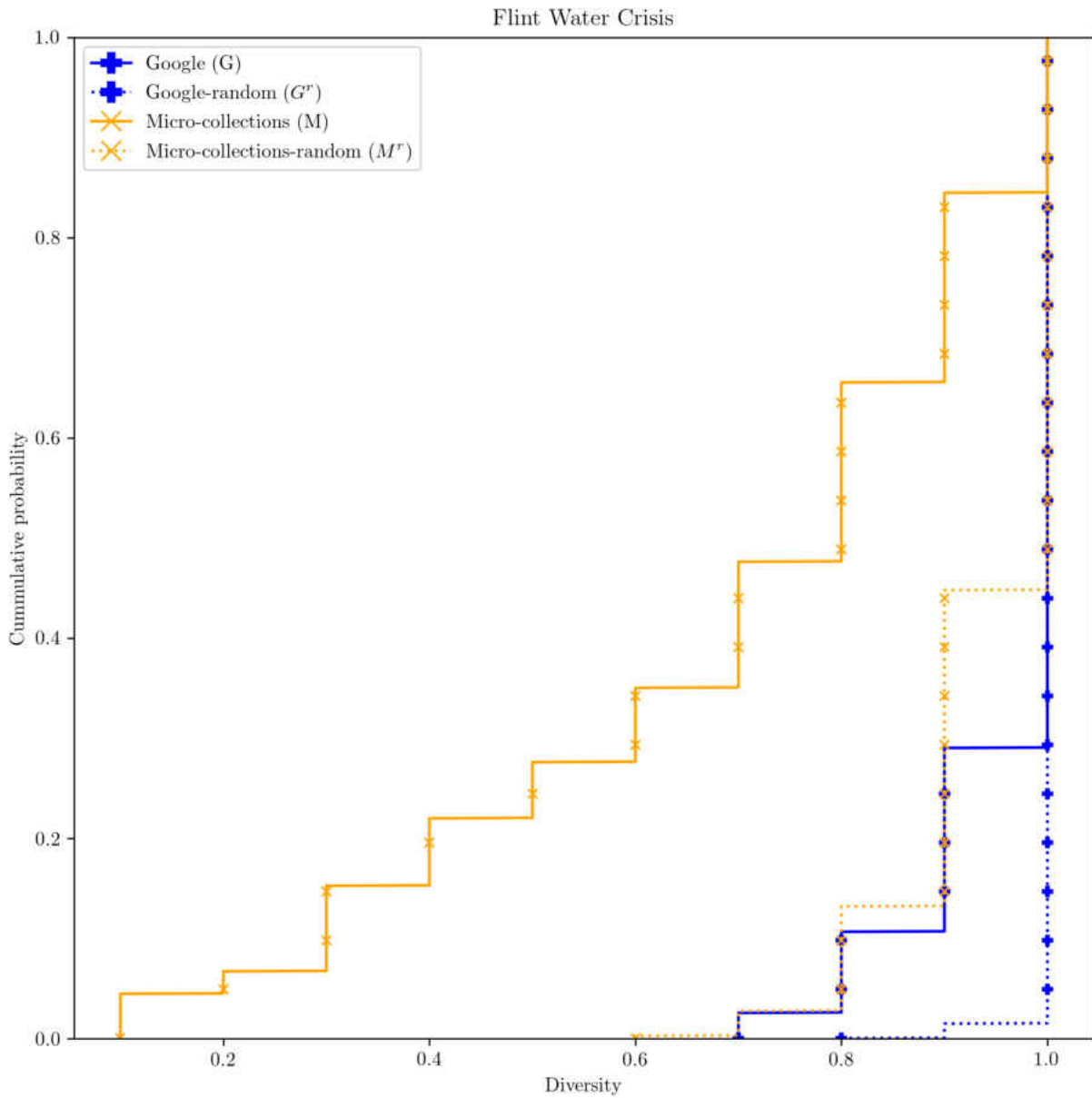


Fig. 64: (Chapter 9.3.4, Flint Water Crisis, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

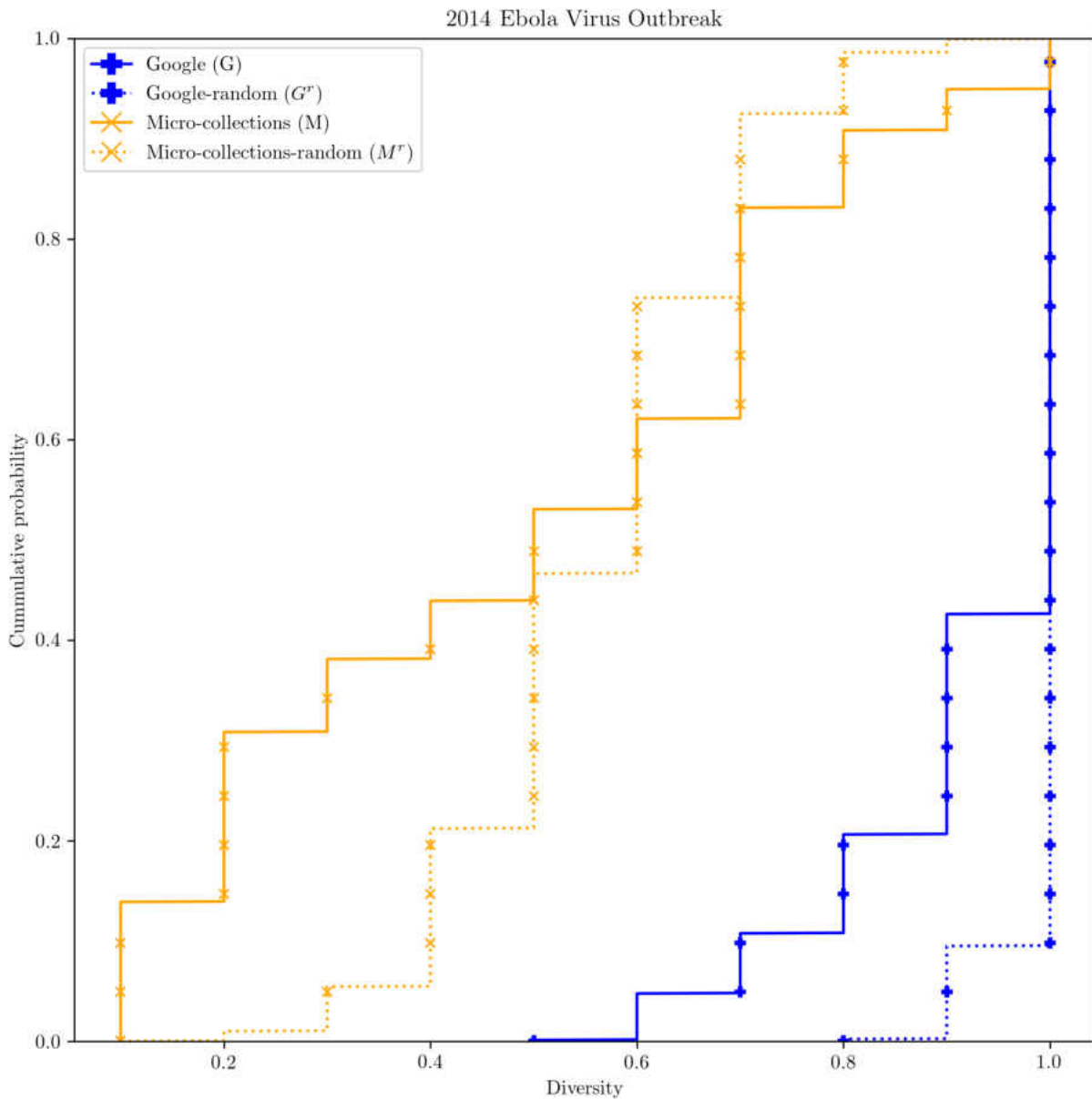


Fig. 65: (Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

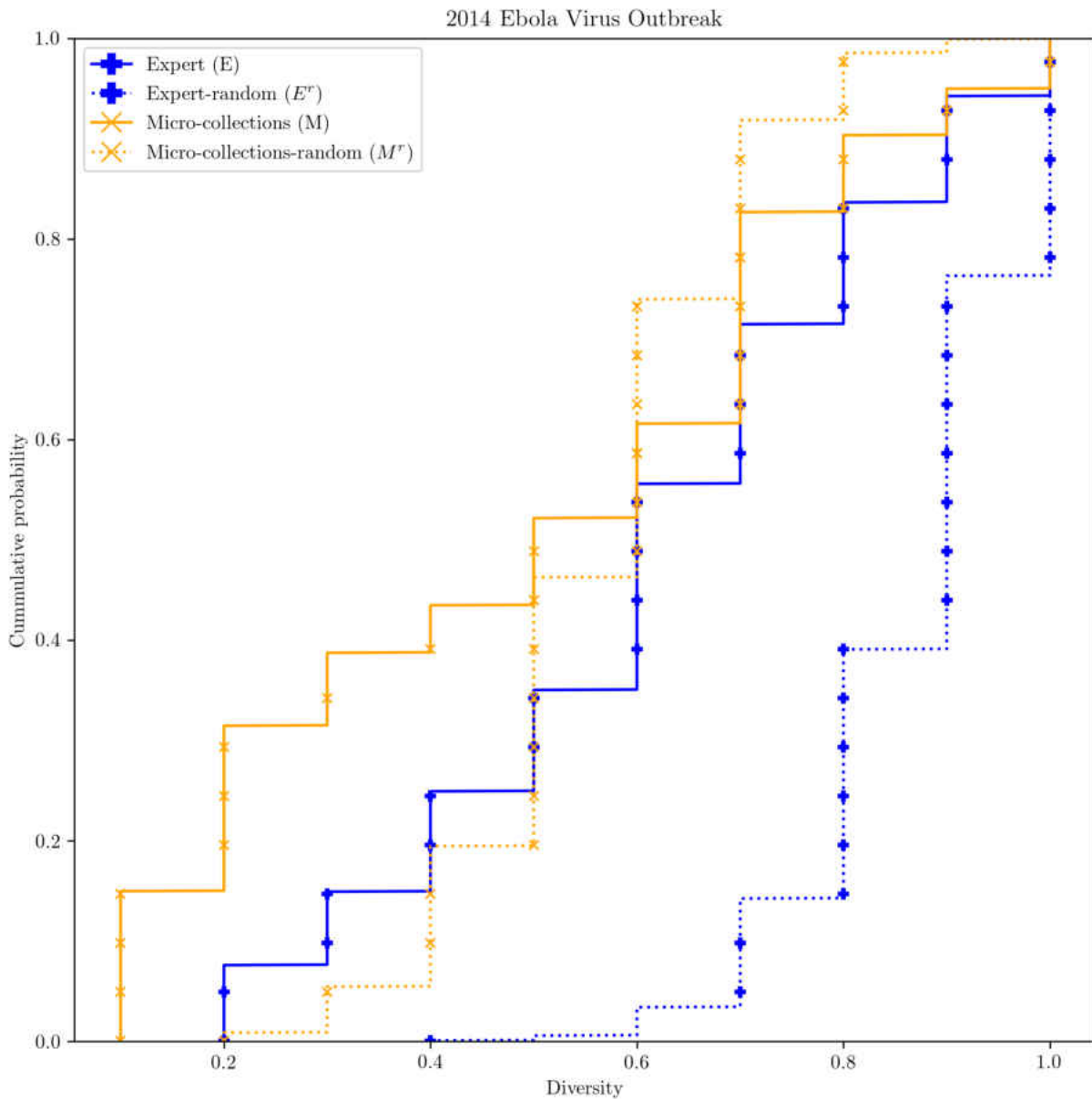


Fig. 66: (Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 44: Empirical Cumulative Distribution Function (ECDF) of the diversity d_u of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

APPENDIX G

**EVALUATION RESULTS: EMPIRICAL CUMULATIVE
DISTRIBUTION FUNCTION OF THE DIVERSITY (D_C - SIZE
CHANGE AFTER COMPRESSION) OF REFERENCE AND
MICRO-COLLECTION SEEDS**

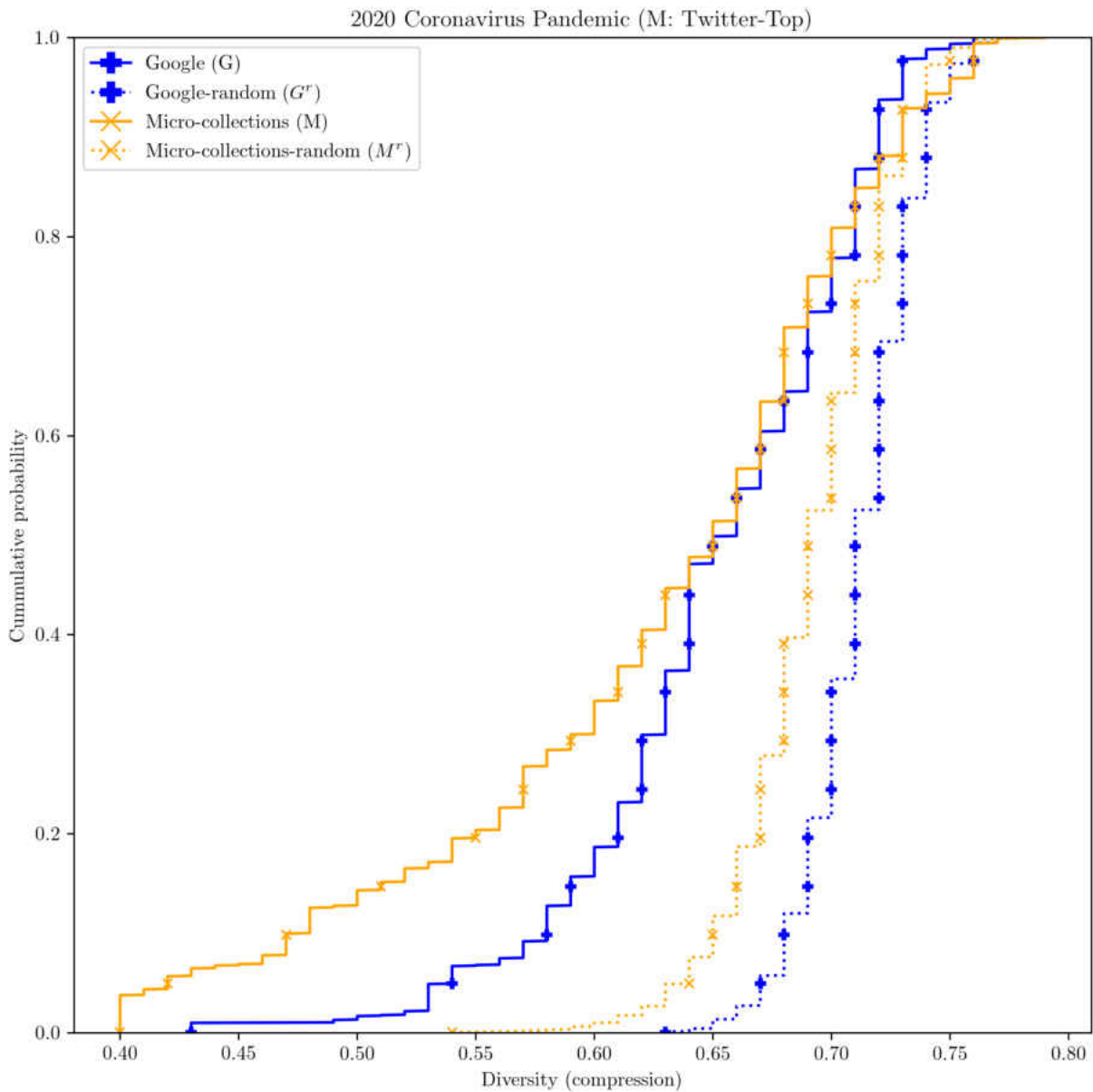


Fig. 67: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

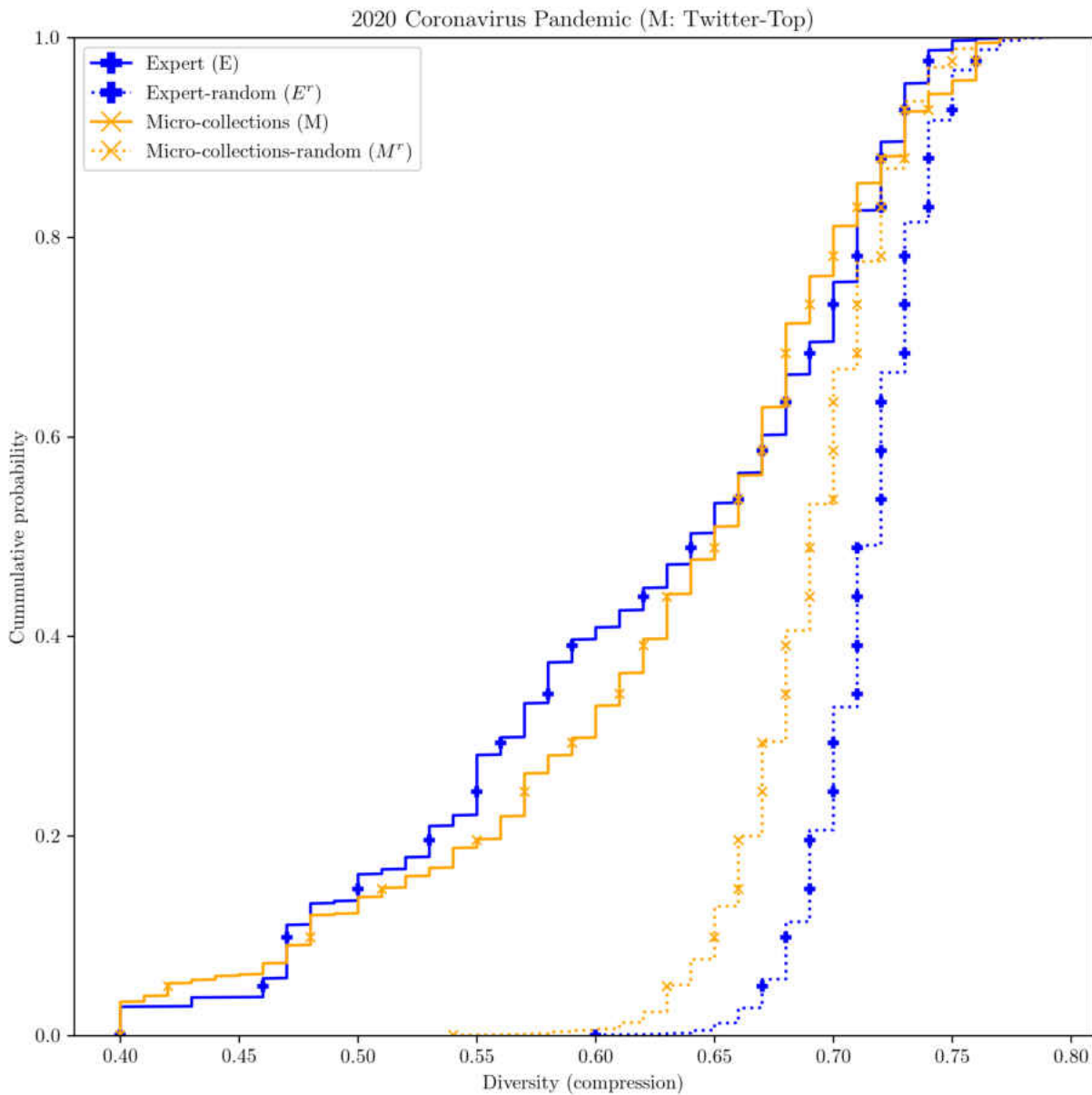


Fig. 68: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

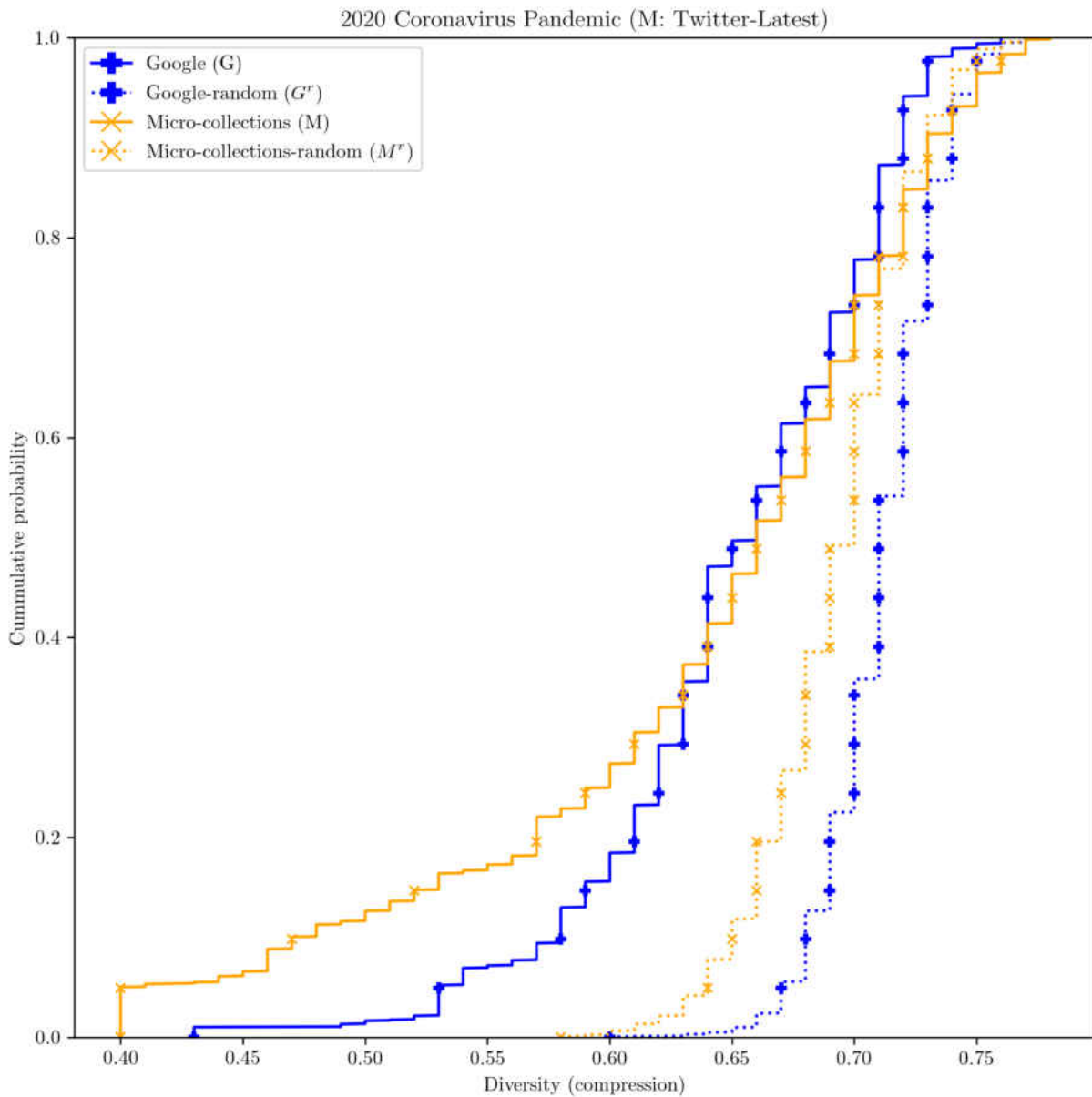


Fig. 69: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

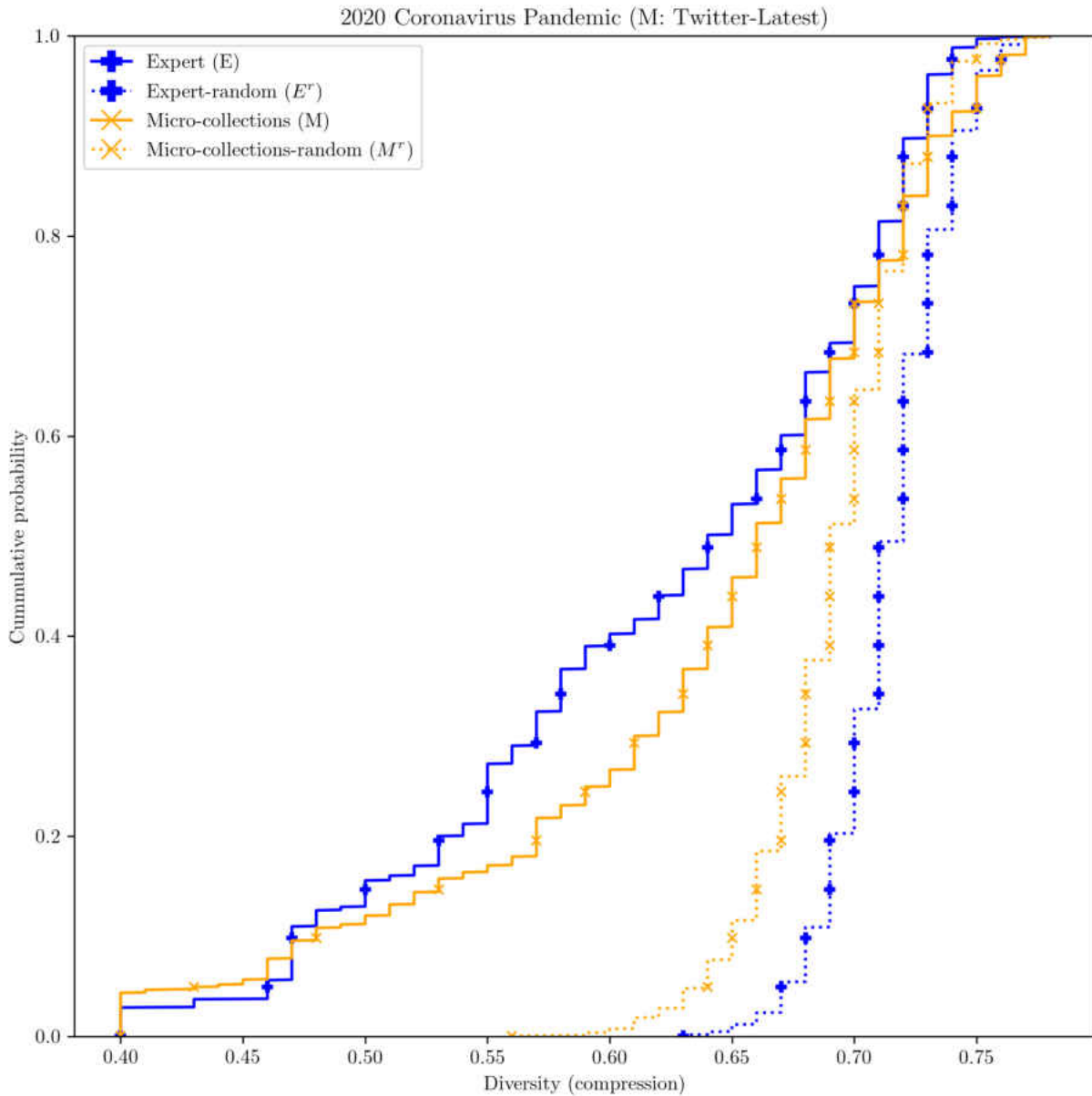


Fig. 70: (Chapter 9.3.4, 2020 Coronavirus Pandemic-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

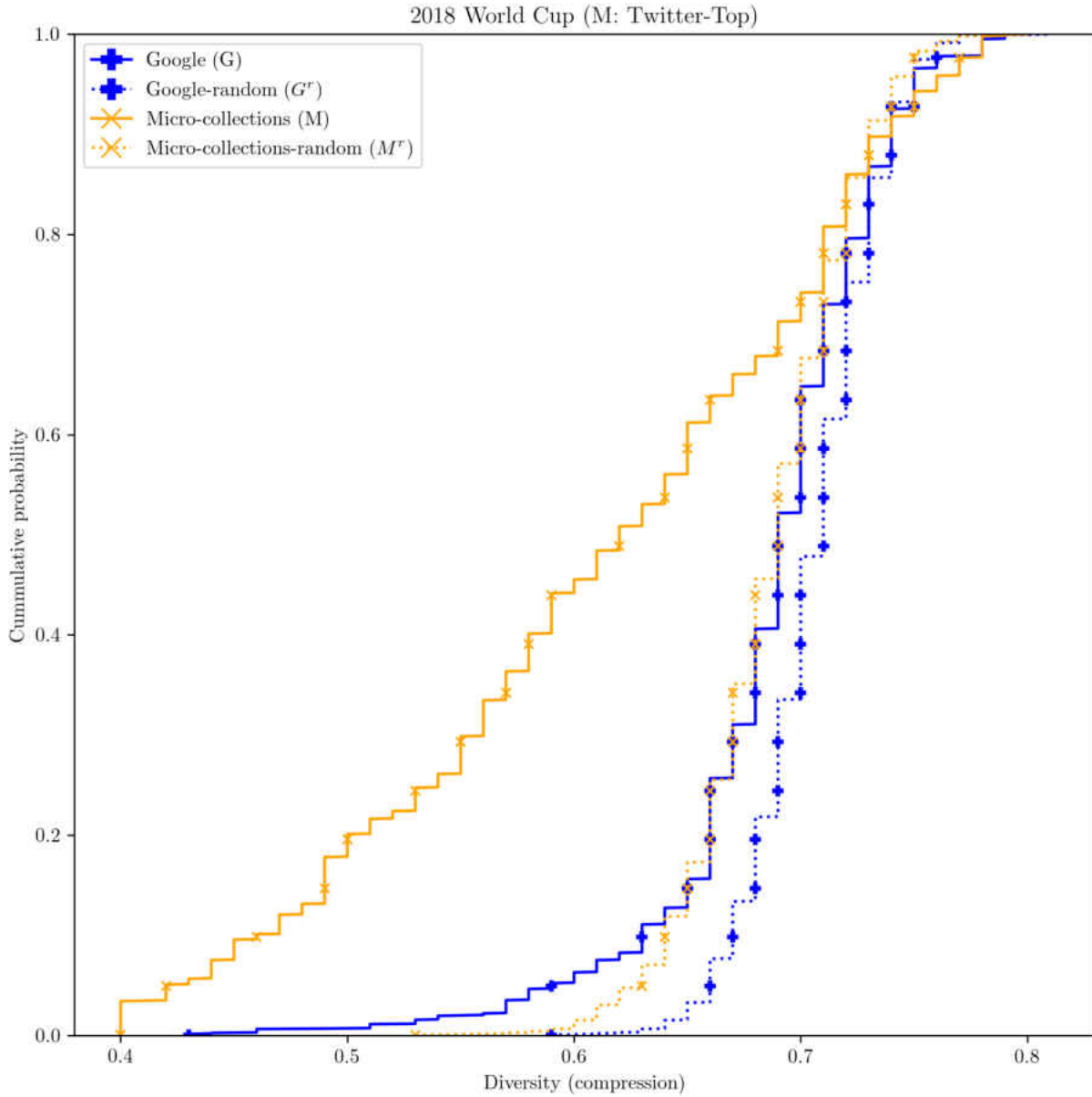


Fig. 71: (Chapter 9.3.4, 2018 World Cup-Top, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

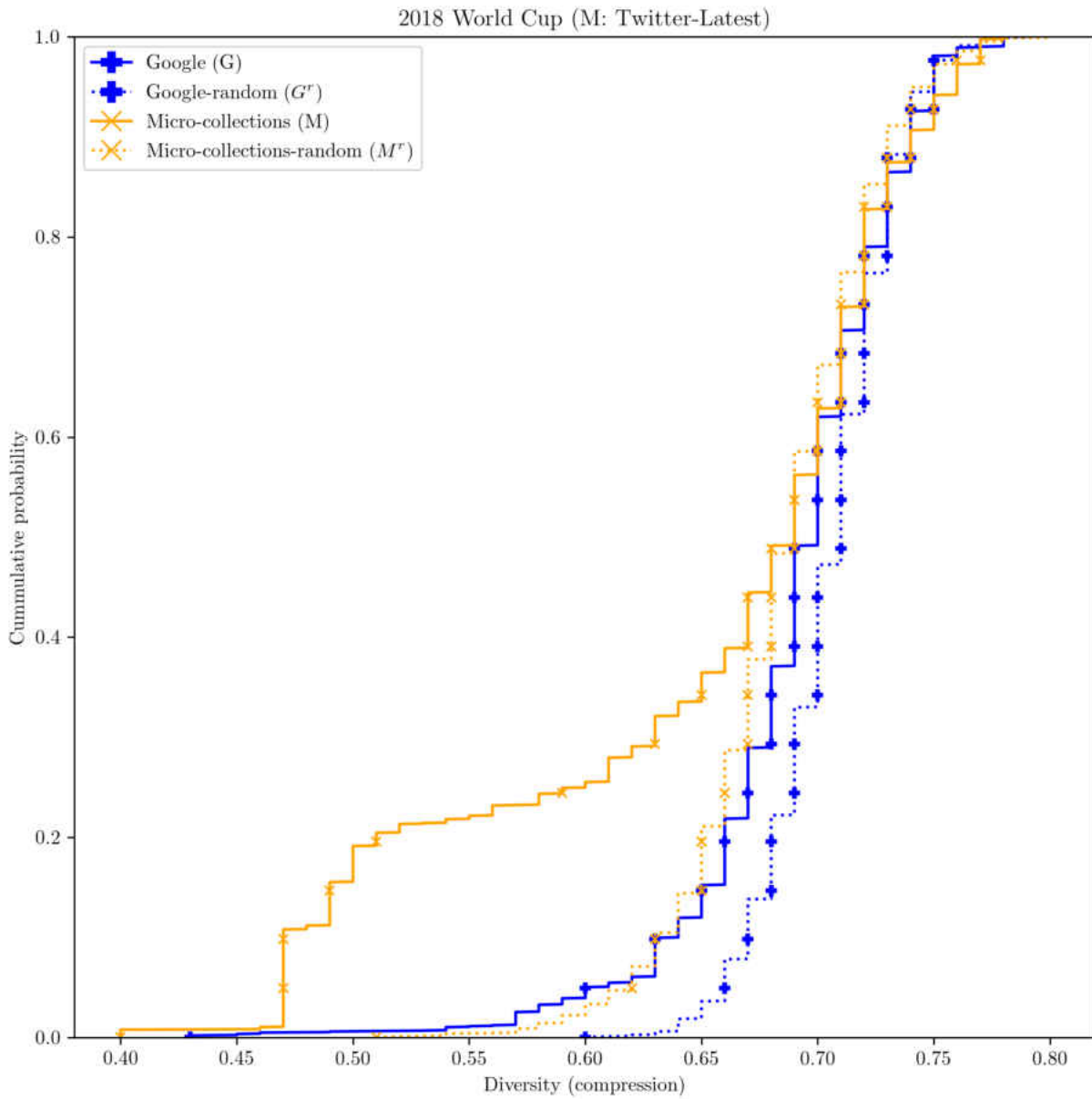


Fig. 72: (Chapter 9.3.4, 2018 World Cup-Latest, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

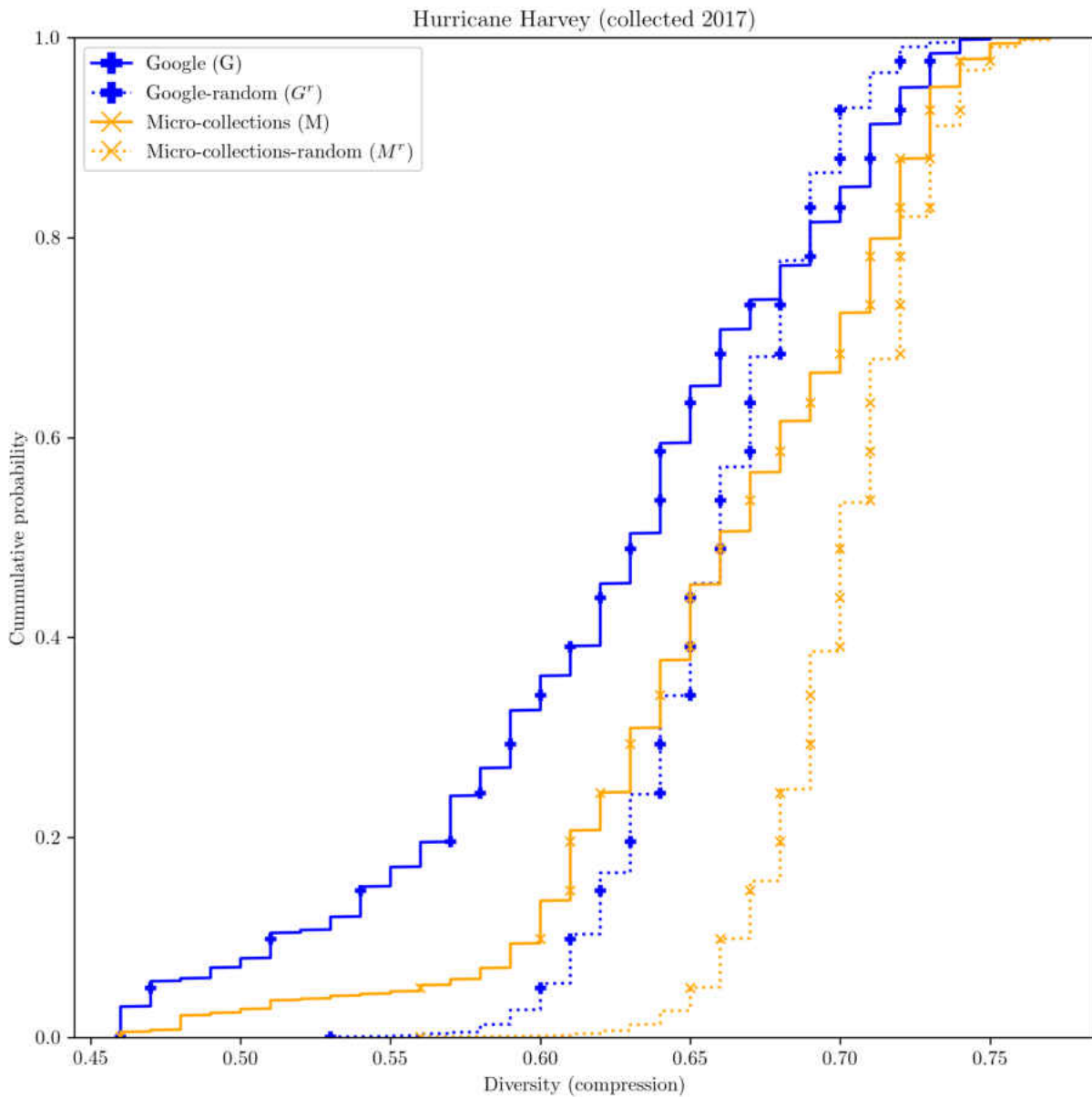


Fig. 73: (Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

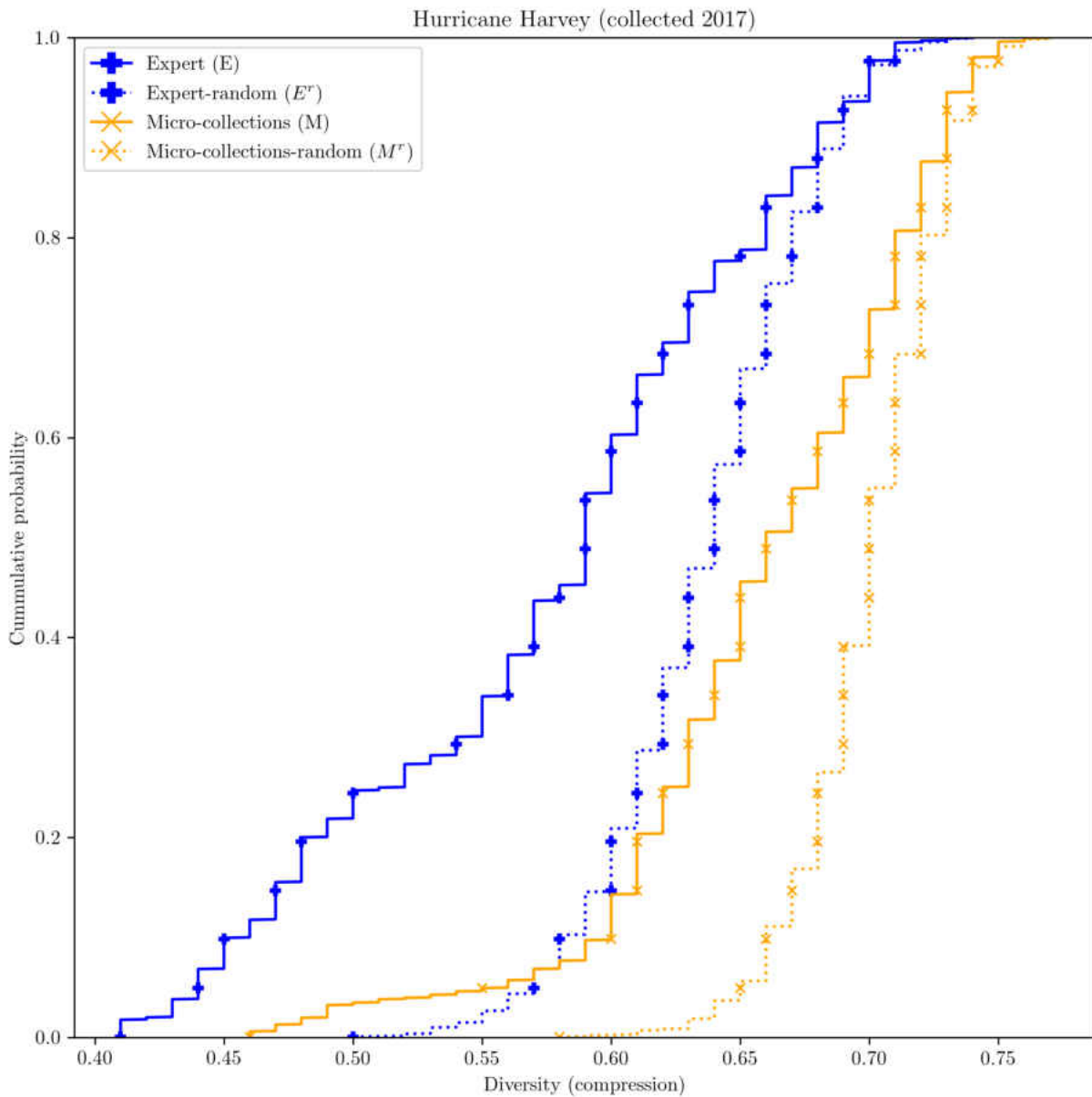


Fig. 74: (Chapter 9.3.4, Hurricane Harvey (collected 2017), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

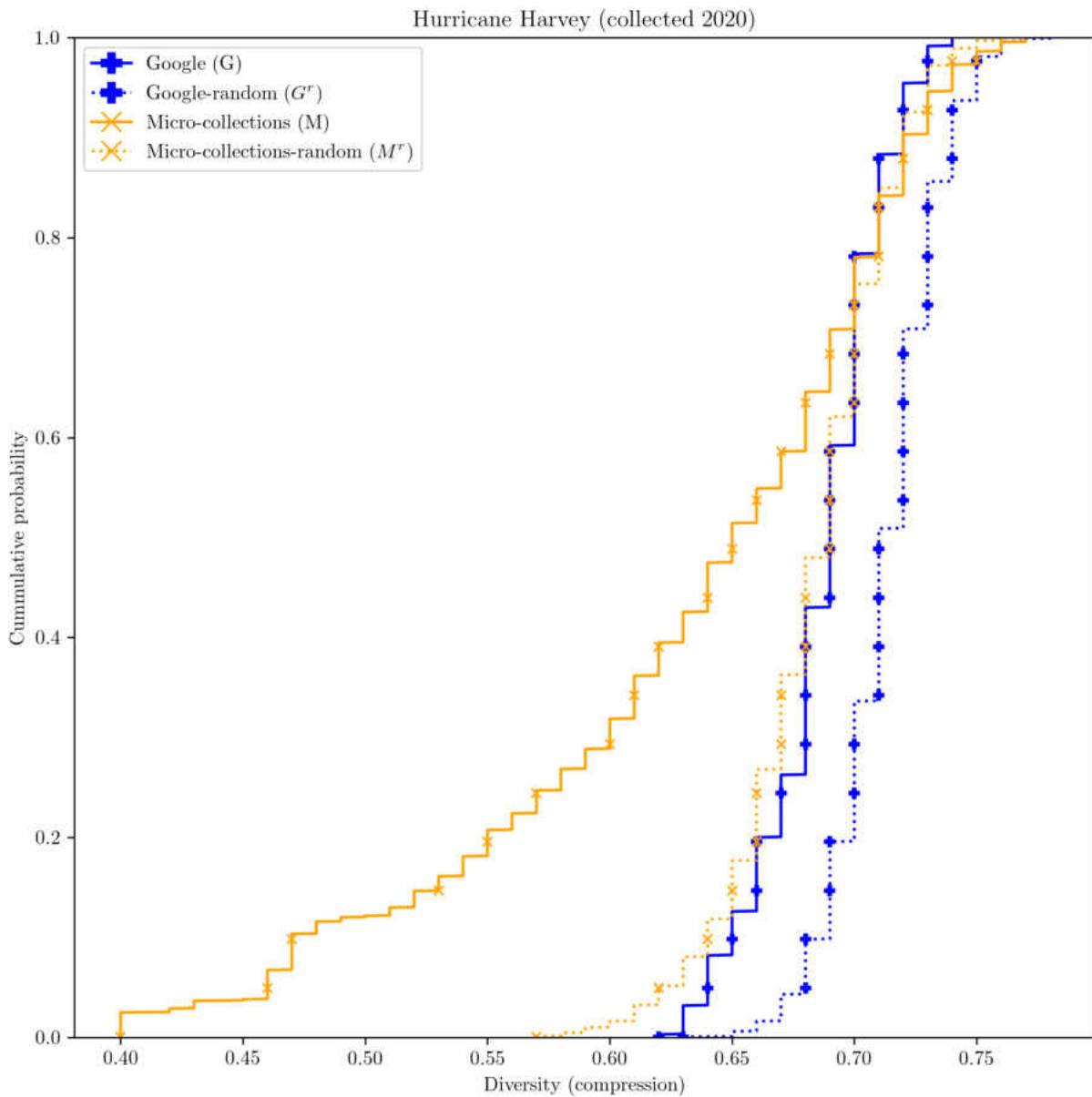


Fig. 75: (Chapter 9.3.4, Hurricane Harvey (collected 2020), Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

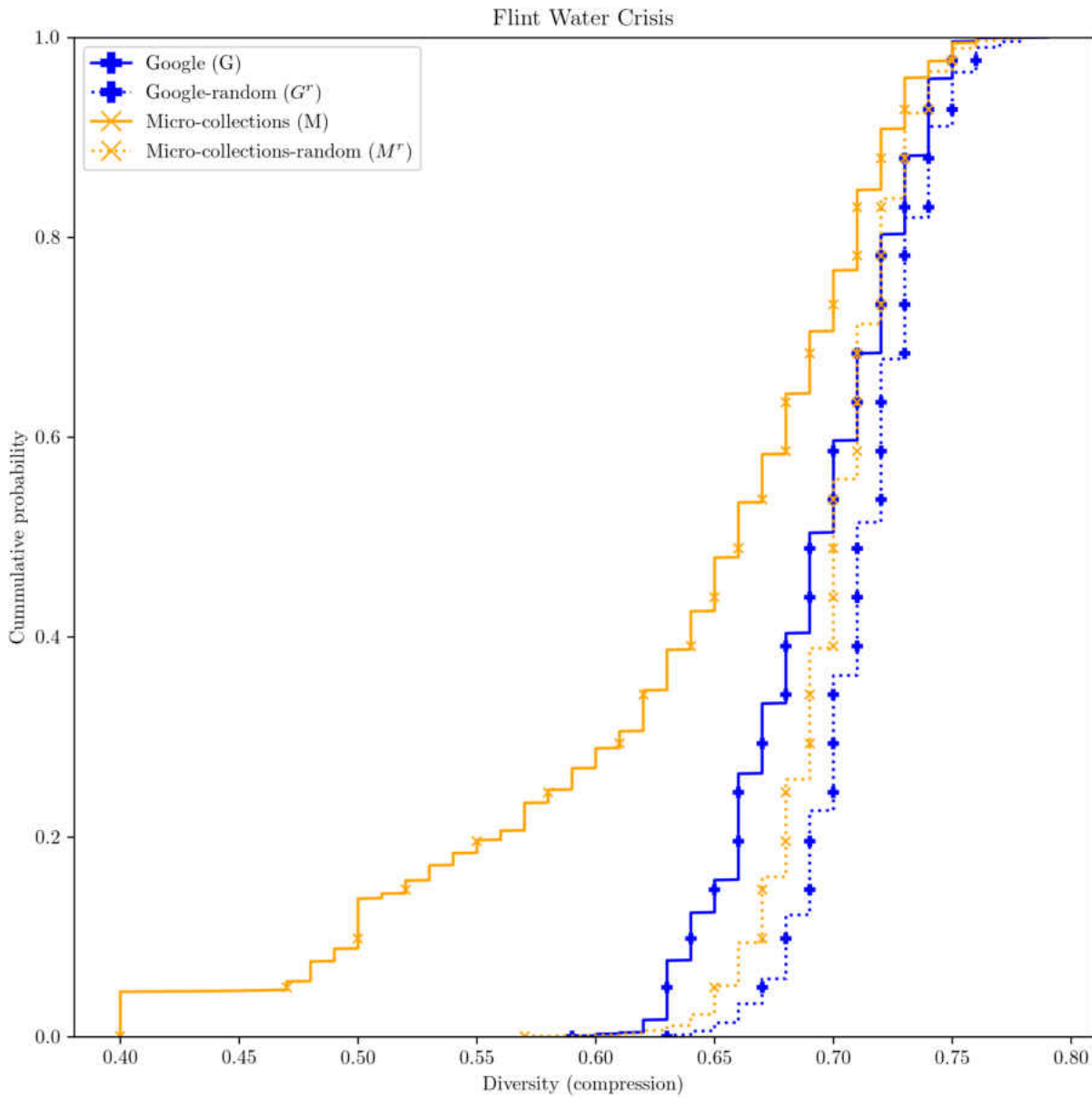


Fig. 76: (Chapter 9.3.4, Flint Water Crisis, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

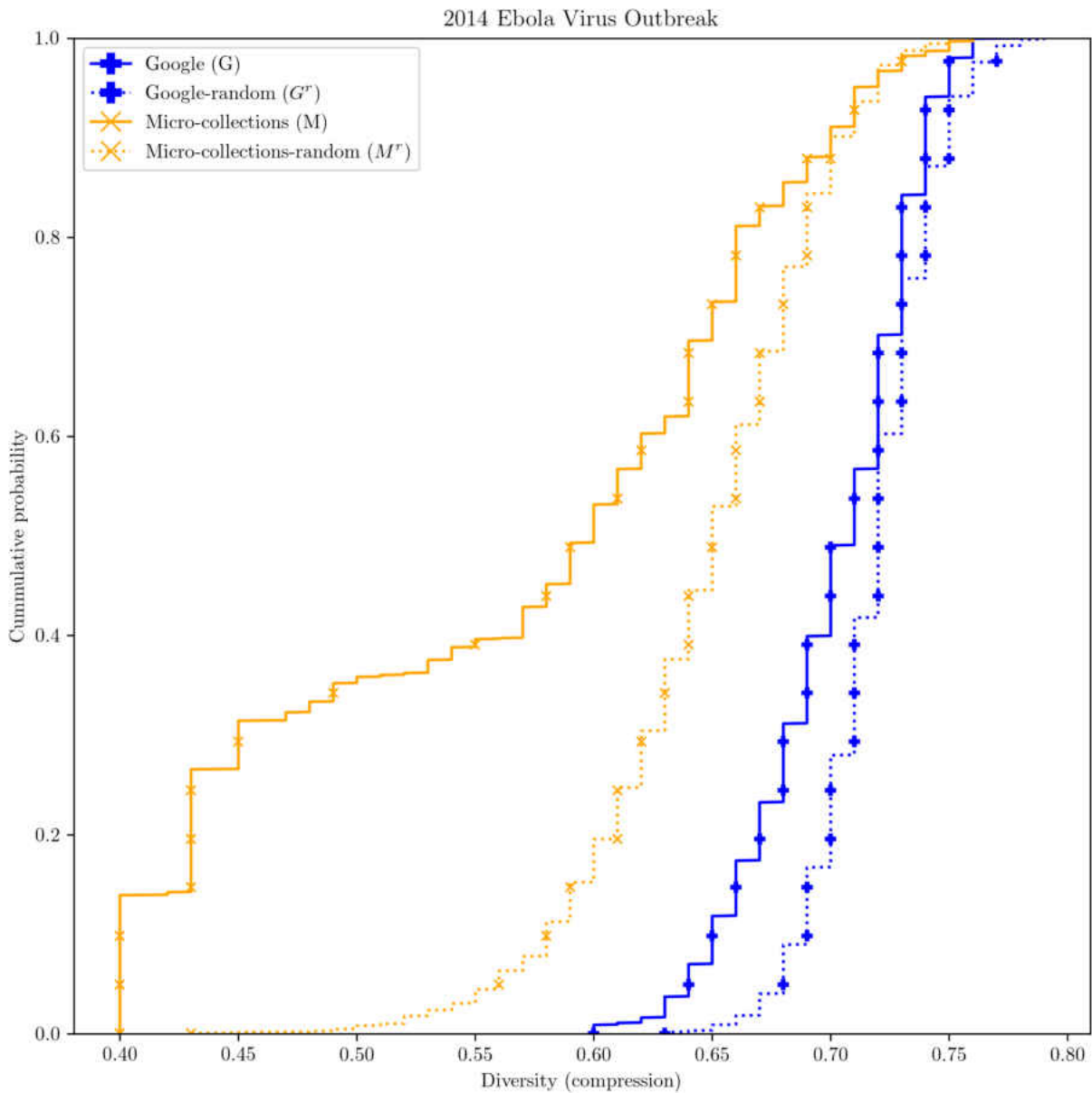


Fig. 77: (Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Google - G and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

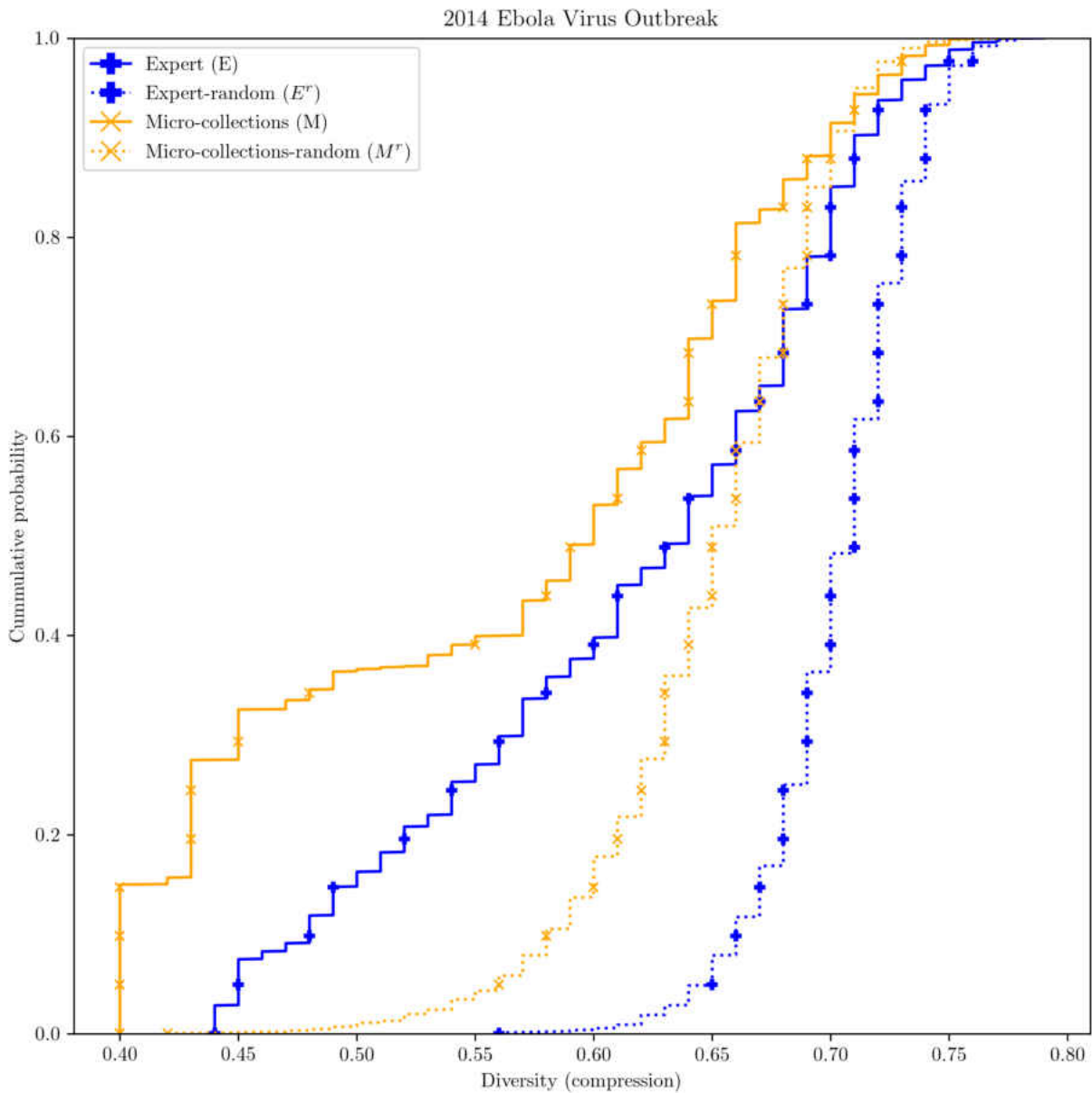


Fig. 78: (Chapter 9.3.4, 2014 Ebola Virus Outbreak, Supplementary visualization of Table 45: Empirical Cumulative Distribution Function (ECDF) of the diversity d_c of seeds of Expert - E and Micro-collections - M seeds. The diversity of seeds selected without QP scores have the r -superscript.

VITA

Alexander C. Nwala
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
e-mail: alexandernwala@gmail.com

Education

Doctor of Philosophy in Computer Science (2020)
Old Dominion University, Norfolk, Virginia USA
Dissertation: *Bootstrapping Web Archive Collections from Micro-collections in Social Media*

Master of Science in Computer Science (2014)
Old Dominion University, Norfolk, Virginia USA
Thesis: *Generating Combinatorial Objects - A New Perspective*

Bachelor of Science in Computer Science (2011)
Elizabeth City State University, Elizabeth City, North Carolina USA

Publications

An updated list of publications is available at <https://scholar.google.com/citations?user=LqrUey4AAAAJ&hl=en>