Trails of Trustworthiness in Real-Time Streams (Extended Summary)

Panagiotis Takis Metaxas*
Harvard University Center for Research on Computation and Society
33 Oxford Str., Cambridge, MA, USA
pmetaxas@seas.harvard.edu

Eni Mustafaraj
Wellesley College
106 Central Str., Wellesley, MA, USA
emustafa@wellesley.edu

ABSTRACT

There is an indisputable need for reliable online information. This need becomes imperative in real-time information channels (R-TICs) that are omnipresent in the Social Web these days. R-TICs are online systems that provide instant interaction, commenting and notification (e.g., Twitter, Facebook, Google+, etc.) While helping us decrease our time and effort to be informed, R-TICs will put new stress to our abilities to act under time pressure in making decisions. Being able to determine the trustworthiness of the information we receive, therefore, will be paramount. How one determines the quality of the information received? Certainly, one needs to be skilled in critical thinking, but technology can also help one act with confidence, by maintaining a network of trusted sources and understanding the reasons why one should trust, or distrust, the information received.

The overall aim of our ongoing research is to lay the foundation of a comprehensive approach to support critical thinking and increase security while maintaining privacy in a trusted cyber-world. Building on the work of other researchers, as well as on the success we had in the past with recognizing and uncovering some of the causes of misinformation, we design a system that can maintain *trails of trustworthiness for information propagated through real-time information channels*. When confronted with information that requires fast action, our system will enable its *educated* users to evaluate its *provenance*, its *credibility* and the *independence* of the multiple sources that may provide this information.

Author Keywords

Social Web; Information Reliability; Social Networks; Trustworthiness; Misinformation Propagation; Twitter.

INTRODUCTION

We are in the initial stages of a profound change in the way we are informed, decide and act. We are emerging from a world where knowledge was primarily produced and analyzed by experts with access to distribution channels such as universities, newspapers and the publishing industry, to one in which everyone can be both producer and consumer of information. Technology will play a central role in this new world, presenting opportunities and dangers. Our research aims to help the citizens of this new world understand both opportunities and dangers, and avoid some of the dangers such as the risks of deception and fraud.

Online social network usage is surging and is expected to increase further in the years to come. College students already use Facebook, myspace, Twitter and other social media daily to be informed about the news at rates far greater than the rates of the established news organizations [33]. Even before the recent surge in the use of online social networks as news sources, an increasing percentage of Americans were using social media to be informed on many financial, medical, religious and political issues. In particular, in 2008, the percentage of Americans that used search engines to be informed reached 83%, while 65% of those from age 18-24 also used an online social network [32, 31].

However, search engines and social networks can be gamed so that they propagate misinformation. For example, the so-called "web spammers" have a track record of forcing their own unreliable content in their top-10 results of search engines by gaming their ranking methods [17, 12, 3, 5, 10, 1, 38, 30, 37] and thrive on Twitter and Facebook [9, 6, 21].

The overall aim of our ongoing research is to lay the foundation of a comprehensive approach to support critical thinking and increase security while maintaining privacy in a trusted cyber-world. When confronted with information that requires fast action, our system will enable its *educated* users to evaluate its *provenance*, its *credibility* and the *independence* of the multiple sources that may provide this information.

What we propose is very ambitious, but we are optimistic due to the work that several researchers, including ourselves, have done in studying the problem of untrustworthy online information in the past. We cite a long, though probably incomplete, bibliography to support our claim. [21, 22] shows that there is a very close relationship between Web Spam in cyberspace and Propaganda in the Society. While propagandists try to alter our personal network of social trust, spammers try to alter the Web Graph of search engines. In fact, one can understand and even predict the tricks employed by

*On leave from Wellesley College

Web spammers by studying the tricks that propagandists employ in real life. Among several examples, we show that the propagandists' *word-games* technique (associating an entity with good or bad words) corresponds to keyword-stuffing for spammers; the *bandwagon* technique ("jump on the bandwagon" since everybody else does) corresponds to creating link farms; the *card stacking* technique (misusing facts and employing illogical derivations) corresponds to link bombs (aka "Google-bombs"). The similarities in the techniques of manipulation of our personal trust networks and the search engines' web graph are remarkably close.

To prove the strength of this relationship between propagandistic and spamming techniques, in [20] we show that one can, in fact, use anti-propagandistic techniques to discover Web spamming networks. In particular, we demonstrate that when starting from an initial untrustworthy site, backwards propagation of distrust (looking at the graph defined by links pointing to to an untrustworthy site) is a successful approach to finding clusters of spamming, untrustworthy sites. This approach was inspired by the social behavior associated with distrust: in society, recognition of an untrustworthy entity (person, institution, idea, etc) is reason to question the trustworthiness of those who recommend it. Other entities that are found to strongly support untrustworthy entities become less trustworthy themselves. As in society, distrust is also propagated backwards on the Web graph.

In cases where there are high stakes, Web spammers' influence may have important consequences for a whole country. For example, in the 2006 Congressional elections, activists using Google bombs orchestrated an effort to game search engines so that they present information in the search results that was unfavorable to 50 targeted candidates [39]. While this was an operation conducted in the open, spammers prefer to work in secrecy so that their actions are not revealed. So, [25] revealed and documented the first *Twitter bomb*, which tried to influence the Massachusetts special elections, showing how an Iowa-based political group, hiding its affiliation and profile, was able to serve misinformation a day before the election to more than 60,000 Twitter users that were following the elections. Very recently we saw an increase in political cybersquatting [16], a phenomenon we reported in [28]. And even more recently, in [27], we discovered the existence of *Pre-fabricated Twitter factories*, an effort to provide collaborators pre-compiled tweets that will attack members of the Media while avoiding detection of automatic spam algorithms from Twitter.

The spammers' activity is, of course, also bad news for the search engines and online social networks that have spent considerable effort in building their reputations [11]. In fact, one can explain the evolution of the various generations of search engines as their effort to counter web spam [21]. It is a war that the search engines have not won, and not for lack of trying or lack of resources. For example, during the 2008 congressional elections, and again during the recent 2010 congressional elections, Google tried to counter spam attacks by presenting carefully selected results in the search for information related to electoral candidates [24].

While Google's approach may be understandable and legal, it also has serious side-effects because it allows the search engines to play the role of "Big-Brother" of reliable information. However, with this approach, search rankings no longer depend on well established algorithms, tested and respected by the research community. Instead, ranking results are decided by a small number of some company's employees. This problem is not restricted to "organic" (algorithmic) search results; it extends and intensifies with the presentation of "paid" results, such as advertisements, as well as real-time search results [28, 29].

REAL-TIME INFORMATION CHANNELS

Of particular interest to our research is the reliability of information that is propagated via real-time information channels (R-TICs), such as the instant interaction, commenting and notification systems that every social network is developing these days (Twitter, Facebook, Google, etc.). Though a relative newcomer in cyber space compared to the Web and the Social Web, R-TICs are expected to play a major role in the near future.

R-TICs will put new stress to our abilities to act under time pressure in making decisions. Consider, for example the situation when an investor receives a message about the looming financial troubles of a company in which she has invested. How sure can she be that this is trustworthy? Should she act upon it, especially when seeing that the stock lost 2% in the last 10 minutes? How can she check for trustworthiness without helping propagate a possible wave of panic? In another situation, a typical voter receives information from apparently the friend of his pastor, that the political candidate he intends to vote for is attacking his religion. Googling about it shows the same message among the top search results. Should he go ahead with his vote or should he switch and inform his relatives? Or, in a more scary situation, consider someone in Jakarta who receives a message apparently sent by the government's disaster advisor, warning about a tsunami moving towards the city. What trustworthiness should he associate with this message? Should he stay put or run? Should he inform his friends and family or look first for verification, and how?

One can come up with lots of examples without using one's imagination, where one needs to act relatively quickly upon receiving some potentially important information. In fact, all three of the examples mentioned above, did happen. The ongoing propagation of fraudulent stock information has a long history [2, 4]. The misinformation that senatorial candidate Martha Coakley was, supposedly, opposing Catholics to be employed in emergency rooms is documented in [25]. The scary misinformation about the fake tsunami was reported in [7]. Financial frauds and missed opportunities abound. Political information and misinformation has been used extensively in the past. Health warnings and exaggerations are reaching us all the time. They are propagated through our social networks and, more recently, through R-TICs. Being able to determine the trustworthiness of the information we receive, therefore, is paramount.

But how one determines the quality of the information one

receives? Certainly, one needs to be skilled in critical thinking and to have reliable sources. However, technology can also help one act with more confidence, by maintaining a network of trusted sources and understanding the reasons why one should trust, or distrust, the information received. Designing such an infrastructure is the main objective of our research.

TRAILS OF TRUSTWORTHINESS

The long-term aim of our research has been to lay the foundation of a comprehensive approach to support critical thinking and increase security while maintaining privacy in a trusted cyber-world. Building on the work of many, we propose a system that can maintain trails of trustworthiness for information propagated through real-time information channels. When confronted with information that requires fast action, our system will enable its users to evaluate its *provenance*, its *credibility* and the *independence* of the multiple sources that may provide this information. However, the system will not be fully automated, and it will be mostly useful to an educated user.

Our concept of trustworthiness comes from the epistemology of knowledge[1]. When we believe that some piece of information is trustworthy (e.g., true, or mostly true), we do so for intrinsic and/or extrinsic reasons. *Intrinsic* reasons are those that we acknowledge because they agree with our own prior experience or belief. *Extrinsic* reasons are those that we accept because we trust the conveyor of the information [36]. If we have limited information about the conveyor of information, we look for a combination of independent sources that may support the information we receive (e.g., we employ "triangulation" of the information paths). In the design of our system we aim to automatize as much as possible the process of determining the reasons that support the information we receive.

We define as *trustworthy*, information that is deemed reliable enough (i.e., with some probability) to justify action by the receiver in the future. In other words, trustworthiness is observable through actions.[2]

The overall trustworthiness of the information we receive is determined by a linear combination of (a) the reputation R_Z of the original sender Z, (b) the credibility we associate with the contents of the message itself $C(m)$, and (c) characteristics of the path that the message used to reach us.

To compute the trustworthiness of each message from scratch is clearly a huge task. But the research that has been done so far justifies optimism in creating *a semi-automatic, personalized tool that will help its users make sense of the information they receive*. Clearly, no such system exists right now, but components of our system do exist in some of the popular R-TICs. For a testing and evaluation of our system we plan to use primarily Twitter, but also real-time Google results and Facebook. Even though these systems are not designed so that they can maintain complete trails of trustworthiness, some parts of the design are testable. We hope that our design will persuade current and future R-TICs adopt some of our design aspects that may enhance their services.

As one of the important domains of information trustworthiness, we plan to use messages related to national elections in the US and abroad since this is an area that there is considerable expertise in information verification (e.g., [34]) and the stakes are high. Of course, the problem of reliable information is not restricted to elections. It can affect any area that involves decisions to be made in a short period of time (e.g., crisis response [18]). However, we have chosen to use the elections since we have already a better understanding of the domain and as an important test case to evaluate the effectiveness of our proposed solutions.

List of Challenges

To address the overall research aim, we put forth several projects that will help the citizen evaluate the quality of information they receive. In particular, we propose:

- Establishment of *new metrics* that will help evaluate the trustworthiness of information people receive, especially from real-time sources, which may demand immediate attention and action. We have experience with identifying quality metrics for search results: In [23] we show that coverage of a wider range of opinions, along with independence of results' provenance, can enhance the quality of organic search results. We plan to extend this work in the area of real-time information so that it does not rely on post-processing procedures that evaluate quality, but on real-time algorithms that maintain a trail of trustworthiness for every piece of information the user receives.

- Monitor the evolving ways in which information reaches users, in particular citizens near election time. This monitoring will help us be informed of the changes introduced by search engines and social media companies as they try to improve their services. Importantly, it will help us examine the – sometimes – unwanted consequences of newly introduced technologies and propose solutions to resolve those in favor of the end user (e.g., [19, 29, 28, 25, 24, 20]).

- Establish a *personalizable* model that captures the parameters involved in the determination of trustworthiness of information in real-time information channels, such as Twitter, extending the work of measuring quality in more static information channels, and by applying machine learning and data mining algorithms. To implement this task, we will design online algorithms that support the determination of quality via the maintenance of trails of trustworthiness that each piece of information carries with it, either explicitly or implicitly. Of particular importance, is that these algorithms should *help maintain privacy* for the user's trusting network.

[1]While there is no clear answer among philosopher on what knowledge is, and why do we believe what we believe, there is a general agreement that critical thinking is one of the better tools that we have to make sense of what we learn.

[2]Note that this is different than the concept of "trust" between users that has been used elsewhere (e.g., [8]).

Even though maintaining trails is a very difficult task to be carried out on the whole Web, it is solvable on a well-designed R-TIC which keeps track of the creation time and propagation path of each message, as well as of its users.

- Design algorithms that can *detect attacks* on R-TICs. For example we can automatically detect bursts of activity related to a subject, source, or non-independent sources. We have already made progress in this area. Recently, we advised and provided data to a group of researchers at Indiana University to help them implement "truthy" [19], a site that monitors bursty activity on Twitter. [3]

  We plan to advance, fine-tune and automate this process. In particular, we will develop algorithms that calculate the trust in an information trail based on a score that is affected by the influence and trustworthiness of the informants.

Most of the ideas we present above can be implemented through plug-ins for browsers and client apps that maintain the trust network of their owner, so that he/she does not have to do the cumbersome part of evidence-finding. Next we discuss some details on the tentative implementation.

CONCLUSIONS

In this paper we have described the seeds of a major project we are currently undertaking. The overall aim of this project is to create the technical and educational components of a comprehensive system that will empower the users of the Social Web. Our system will optimally be used by an educated user and will be able to help her stay informed and make decisions on important issues that she cares about, in a short period of time.

On the technical side, we observe that in the past few years there has been a lot of work in studying social networks with tools from graph theory, artificial intelligence, machine learning, statistical methods, interface design, and human-computer interaction, that justifies our optimism. We are building on the considerable success of many researchers, including ourselves.

We expect our project to have significant impact on how people use and understand social networks. According to a recent Pew report, the number of people who use regularly micro-blogging and Twitter doubled in the last few months. Having an R-TIC client tuned to its particular user will enhance peoples understanding of the benefits and risks of using social networks on a daily basis. We aim to have users think in terms of trustworthiness and independence of the information they receive in their daily lives.

Of course, people are able to maintain inconsistencies in their belief system and are susceptible to psychological and societal biases (e.g., confirmation bias [26]). Our system will provide evidence of such inconsistencies for those who are willing to examine them ("This seems false though I have trusted the conveyor of the information in the past"). It is not aimed to reveal automatically the truth behind every piece of information that comes to the user, but to help those what are willing to search for it, by automatizing only those actions that can help them determine what to trust.

We should mention that in a month from this writing, Ushahidi [35], a crowd sourcing platform aiming to help in the coordination of humanitarian support, plans to release Swift River, a platform that "enables the filtering and verification of real-time data from channels like Twitter, SMS, Email and RSS feeds". Several of the features of Swift River seem similar to what we propose [40], though a major difference appears to be that our design is personalization at the individual user level.

[3]There has been a lot of public interest in this work, as seen by extensive media coverage (e.g., in The Atlantic [14], the Chronicle of Higher Education [13] and Technology Review [15], to name but a few).

ACKNOWLEDGMENTS

This research was partially supported by NSF grant CNS-1117693. The authors would like to thank Catherine Wearing, Ethan Zuckerman and the Fellows of the Berkman Institute for valuable discussions on the epistemology of knowledge.

REFERENCES

1. Benczúr, A., Csalogány, K., Sarlós, T., and Uher, M. Spam Rank – Fully automatic link spam detection. In *Proceedings of the AIRWeb Workshop* (May 2005).
2. Berenson, A. On hair-trigger wall street, a stock plunges on fake news. New York Times, Aug. 26 2000.
3. Bianchini, M., Gori, M., and Scarselli, F. PageRank and web communities. In *Web Intelligence Conference 2003* (Oct. 2003).
4. Corey, T. S. Catching on-line traders in a web of lies: The perils of internet stock fraud. Ford Marrin Esposito, Witmeyer & Glesser, LLP, `http://www.fmew.com/archive/lies/`, May 2001.
5. Fetterly, D., Manasse, M., and Najork, M. Spam, damn spam, and statistics. In *WebDB2004* (June 2004).
6. Gayo-Avello, D., and Brenes, D. Overcoming spammers in twitter a tale of five algorithms. In *Proceedings of the CERI 2010 Conference* (Madrid, Spain, June 15-16 2010).
7. Globe, T. J. Government disaster advisors twitter hacked, used to send tsunami warning. `http://bit.ly/fODgms`, Last retrieved on Nov., 11, 2010.
8. Goldberg, J. *Computing and Applying Trust in Web-Based Social Networks*. University of Maryland, College Park, Ph.D. thesis, 2005.
9. Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: The underground on 140 characters or less. In *Proceedings of the ACM CCS'10 Conference* (2010).
10. Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. Combating web spam with TrustRank. In *VLDB 2004* (Aug. 2004).

11. Hansell, S. Google keeps tweaking its search engine. New York Times, Jun. 3 2007.
12. Henzinger, M. R., Motwani, R., and Silverstein, C. Challenges in web search engines. *SIGIR Forum 36*, 2 (2002), 11–22.
13. Kaya, T. Separating the truth from the truthy. The Chronicle of Higher Education, `http://bit.ly/oljiUx`, Last retrieved on Nov., 9, 2010.
14. Keller, J. When campaigns manipulate social media. The Atlantic, `http://bit.ly/nzGTS8`, Last retrieved on Nov. 11, 2010.
15. Kleiner, K. Bogus grass-roots politics on twitter. Technology Review, `http://bit.ly/zT00Rz`, Last retrieved on Nov., 9, 2010.
16. Lacey, M. Clicking candidate.com, landing at opponent.com. New York Times, Sept. 14 2010.
17. Lynch, C. A. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *J. Am. Soc. Inf. Sci. Technol. 52*, 1 (2001), 12–17.
18. Meier, P. How to verify social media content: Some tips and tricks on information forensics. iRevolution Blog, `http://bit.ly/yHYoLA`, Last retrieved on Nov., 11, 2011.
19. Menczer, F., Flammini, A., Bollen, J., and Vespignani, A. Truthy. http://truthy.indiana.edu/, Last retrieved on Nov., 9, 2010.
20. Metaxas, P. T. Enhancing information reliability through backwards propagation of distrust. *International Journal of Advances in Security 2*, 2–3 (August 2009), 214 – 225.
21. Metaxas, P. T. Web spam, social propaganda and the evolution of search engine rankings. *LNBIP 45* (2010), 170–182.
22. Metaxas, P. T., and Destefano, J. Web spam, propaganda and trust. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web* (Chiba, Japan, May 2005).
23. Metaxas, P. T., Ivanova, L., and Mustafaraj, E. New quality metrics for web search results. *LNBIP 18* (August 2009), 280–294.
24. Metaxas, P. T., and Mustafaraj, E. The battle for the 2008 us congressional elections on the web. In *Proceedings of the Web Science 2009 Conference* (Athens, Greece, March 2009).
25. Metaxas, P. T., and Mustafaraj, E. From obscurity to prominence in minutes: Political speech and real-time search. In *Proceedings of the Web Science 2010 Conference* (Raleigh, NC, April 2010).
26. Mooney, C. The science of why we don't believe science. Mother Jones, http://bit.ly/xOT4dZ, Last retrieved on Feb., 7, 2012.
27. Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. T. Vocal minority versus silent majority: Discovering the opinions of the long-tail. In *Proceedings of the Third IEEE International Conference on Social Computing* (2011).
28. Mustafaraj, E., and Metaxas, P. T. Sponsored search for political campaigning during the 2008 us elections. In *Proceedings of the 2009 SIGIR-IRA Conference* (Boston, MA, July 2009).
29. Mustafaraj, E., Metaxas, P. T., and Grevet, C. The use of online videos in the 2008 us congressional elections. In *Proceedings of the SocialComm 2009 Conference* (Vancouver, Canada, August 2009).
30. Ntoulas, A., Fetterly, D., Manasse, M., and Najork, M. Detecting spam web pages through content analysis. In *WWW 2006* (May 2006).
31. Pew Foundation. Internet's broader role in campaign 2008 - Social networking and online videos take off. `http://bit.ly/w5rJG9`, January 11 2008.
32. Pew Foundation. Key news audiences now blend online and traditional sources. http://people-press.org/report/?pageid=1354, August 17 2008.
33. Spitzer, R. *Journalism, the Internet, and Online Social Networks: New Patterns of News Consumption in the Wellesley Community and Beyond*. Wellesley College, Honors thesis, 2010.
34. Times, S. Politifact. `http://bit.ly/xiH7TE`, Last retrieved on Nov., 11, 2011.
35. Ushahidi. Technology innovations for humanitarian assistance. http://ushahidi.com/, Last retrieved on Feb. 6, 2012.
36. Vedder, A. Reliability of information: Some distinctions. In *Proceedings of the Computer Ethics - Philosophical Enquiry (CEPE2003) Conference* (2003).
37. Wang, Y.-M., Ma, M., Niu, Y., and Chen, H. Spam double funnel: Connecting web spammers with advertisers. In *WWW 2007* (May 8–12 2007).
38. Wu, B., and Davison, B. Identifying link farm spam pages. In *Proceedings of the fourteenth international conference on World Wide Web* (May 2005).
39. Zeller Jr., T. Gaming the search engine, in a political season. New York Times, Nov. 6 2006.
40. Zuckerman, E. Personal communication, Jan. 24, 2012.