In: B. Kennedy, D. Reitter, R. St. Amant (Eds.), *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation*. Ottawa, Canada: BRIMS Society, 2013.

Trust Definitions and Metrics for Social Media Analysis

Eli Stickgold
Corey Lofdahl
Michael Farry
Charles River Analytics
625 Mt. Auburn Street
Cambridge, MA 02138
617-491-3474
estickgold@cra.com, clofdahl@cra.com, mfarry@cra.com

Keywords:
trust theory, trust metrics, social networks, graph theory, automated social media analysis

ABSTRACT: Social media has changed the information landscape for a variety of events including natural disasters, demonstrations, and violent crises. During these events, people use a variety of social media, such as Twitter, to share information with the world. Given the massive amount of data generated, it is difficult to identify the valuable information in a sea of noise. In this study, we focus on a universal contributing factor to information value—trust—which is analyzed in three steps. First, a theory of *trust in information* is developed based on the application of existing literature to this unique and emerging issue. Second, a set of metrics is developed that focus on trusted relationships and behavioral indicators of trustworthiness within social media. Third, these trust metrics are tested on an anonymized data set and their results presented.

ACKNOWLEDGEMENTS: This work was performed under US Navy contract number N00014-12-M-0259. The authors thank Dr. Rebecca Goolsby for her significant technical support and eager engagement on this project. This work was funded in its entirety by the Office for Naval Research (ONR). We also acknowledge the contribution of Professor Adam Henry of the University of Arizona who introduced the authors to the concept of trust in information.

1. Introduction

The proliferation of social media has changed the information landscape for a variety of events including natural disasters, demonstrations, and violent crises. During these events, people use a variety of social media (e.g., Twitter) to share information. Given the massive amount of data that can be generated in a short period of time by social media, it is difficult to identify valuable information in a sea of noise. That analysis is dependent on careful definitions of information value, which may shift according to the type of data considered, who is inspecting the data, and for what purposes the data is being used. To accommodate that broad range of definitions, many current attempts to exploit social media are highly manual, using "mechanical Turk" approaches to synthesize information. For example, the disaster response information sharing service Ushahidi (http://www.ushahidi.com) used a team to manually categorize and store responses to the Haiti earthquake in January 2010. However, reliance on manual synthesis impedes scaling. As a result, approaches like Ushahidi are not well-suited in all situations to determine information value. A more robust solution, one that explicitly models information value in more automated and decision-relevant terms, is clearly required.

This paper examines the notion of *trust* as a way to determine key information within social media using automated tools, which is accomplished in three steps. First, the theory of trust in information is developed, based on the application of existing literature to this unique and emerging issue. Second, a set of metrics is developed that focus on trusted relationships within social media data sets. These metrics focus on the structure of relationships rather than the content of messages passed within social media, which we posit as a more reliable indicator of trust. Third, these trust metrics are then tested on an anonymized Twitter data set from the Occupy Wall Street (OWS) movement. The results of this test indicate that trust relationships can be found within social media using automated tools.

2. Trust Theory

Trust is central to the study of society, groups, and organizations because it provides the fundamental and basic motivation behind people coming together to cooperate for a common purpose (Arrow 1974; Poteete et al. 2010, 226-227). Concepts about trust however are becoming more complicated and complex as people communicate and form networks through social media (Henry and Dietz 2011). Traditional trust studies focused on cases from small, isolated, and institutional settings (Hahn et al. 2006; Ruttan 2006) or drew conclusions from controlled experiments (Bouma et al. 2008; Vollan 2008). Such studies though are insufficient to address real-world social interactions (Sabatier 1999), let alone the increased complexity of modern social media. This section extends traditional definitions of trust in three parts. First, traditional trust concepts are discussed as, "trust in action." Second, the complexities presented by social media are discussed. Third, a set of "trust in information" concepts are offered to address the analytic challenges posed by modern networks and social media.

Traditional notions of trust are based upon small group settings in which repeated face-to-face interactions are the norm (Siegrist et al. 2007; Renn 2008:222–230). Through multiple, reciprocal, and long-term social interactions, trust among the small group members is built up over time. *Trust in action* is rooted in such settings and is based on a set of related behaviors, three of which are long-term and one short-term (Coleman 1990; Sztompka 1999). First, people have varying levels of willingness to take trusting actions, in which trust is defined as risking something of value based on the actions of others. Second, people exhibit varying levels of aversion to betrayal—that is, situations in which something of value is lost based on the actions of others. Third, people also exhibit varying degrees of altruism as defined by concern for the well-being of others. Fourth, people calculate and modify betrayal odds based on direct calculation, a short-term behavior that uses currently available information. So the first three personal behaviors—willingness to trust, aversion to betrayal, and altruism—are long-term characteristics that are resistant to short-term changes. The fourth, calculation of betrayal odds, is more short-term and thus more modifiable. Together, these four behaviors form the analytic baseline of trust in action.

Trust in action—based as it is in small group, face-to-face settings—needs to be modified to address real-world social processes that encompass larger numbers of people and more complex dynamics (Sabatier 1999). The recent growth of social media provides the potential for larger empirical studies of trust to be performed (Adali et al. 2010), and the increased importance of these social media provides the incentive to undertake such studies. Initial research indicates several important and emerging system features. Trust studies in large social networks exhibit *polycentrism*, the presence of multiple decision centers or groups, each with its own rules and relationships (Young 2002). As trust studies transition from small group settings to social media, polycentrism will prove an increasingly salient system feature. The empirical data available from social media also provides the potential to yield insights into the ways that trust evolves over time and leads to collaboration among seemingly divergent groups (Fehr 2009). This evolving conception of trust can also lead to group formation that reinforces polycentrism, as people who have collaborated in the past tend to collaborate in the future. Separate groups with different beliefs and internal structures may also lead to different information uptake and processing dynamics through "confirmation bias," in which people tend to believe information that confirms what they already know (Munro and Ditto 1997). Finally, trust in social media will be based on the information available through social media rather than real-world actions observed in a face-to-face setting, which will have social consequences. These questions pose research opportunities for social media and the networks that emerge from them because while much information is available, that information needs to be organized, fitted, and understood in order to provide trustable social media insights.

Trust in information provides a framework for trust studies in social media by considering both behavioral indicators (reflected through communications network structure) and information content. Two key features of network structure are important for trust in information. The first network feature concerns two-way relationships such as transitivity and reciprocity that represent interactions with the potential for trust in social media (Fehr and Gächter 2000; Rasmusen 2001). The second network feature concerns reputation or centrality, which indicates status or leadership within the social network. Such status within a network can in turn lead to preferential attachment dynamics, in which those with many social connections or high status then tend to exhibit even more connections over time (Snijders et al.

2006). Besides network position though, the content of the messages that pass among the members of the network also need to be considered because it is through messages that individuals ascertain whether others are similar to themselves. Similarity between those in a network, i.e. *homophily*, plays an important role in social media dynamics because people tend to trust those who exhibit similar traits, thoughts, and opinions (McPherson et al. 2001). This process of self-selection helps to drive group formation and polycentrism, so determining just how those in social media determine similarity and how those homophily-driven processes play out is an important feature of trust.

To conclude, trust in information leverages trust in action by focusing on network position to determine transitivity, reciprocity, reputation, and centrality. Analyzing message content to determine similarity and homophily among individuals is recognized as important but is not addressed herein. Our initial attempts to implement these ideas are discussed in the following section.

3. Metric Development

We now present structure-based metrics that support the identification of trustworthy individuals. As discussed above, our concept of trust in information considers network structural elements rather than message content, which is outside the scope of this study. Before delving into the nature of these metrics, we note that these metrics provide critical cues about trustworthiness, but they are not the only indicators of trustworthiness. For each analytical application, the theory of trust in information naturally takes a different quantitative form. We expect that each application would use these metrics in a significant way, but that trustworthiness assessments would never be based wholly on these metrics.

We began by formalizing a conceptual hypothesis: that a node (or social media user) that served as a source of information to high-profile and widely-distributed users (e.g., news sources, celebrities) could provide high-value and potentially trustworthy information. This hypothesis leads to two graph-theoretic definitions. We define *first-order influence* as simply the degree to which a user receives attention from others in a social network. First-order influencers are not necessarily inherently trustworthy—news sources may be biased or untrustworthy in many areas of the world, and the same is true of other central figures such as celebrities. Rather, it is the second concept that allows for the identification of trust relationships. We define *second-order influence* as the degree to which a user *indirectly* receives attention in a social network. Second-order influence is measured by the propagation of a user's ideas into a social network by means other than direct messages. We hypothesized that second-order influencers are more likely to provide trustworthy information, since first-order influencers look to them for local, timely, and detailed knowledge. In our OWS Twitter data set, we had access to the messages, but not to follower data for the users in question. In situations like this, a simple metric like degree centrality performed on a directed unweighted "mention network" in tweets can identify users with high first-order influence. Without resorting to computationally-intensive and unreliable content analysis techniques, we exploited a simple Twitter social convention, retweeting, to track second-order influence. While this specific mechanism is unique to Twitter, the general concept applies to most social media (e.g., resharing on Facebook), email systems (forwarding or quoting), and collaborative data sharing and office productivity software (through mechanisms as simple as cutting and pasting).

In our test data, we considered users to have high second-order influence if they were frequently retweeted by users with high first-order influence. Several metrics, such as degree centrality, betweenness centrality, Katz centrality or the PageRank metric, can identify users with high second-order influence, but our desired property is slightly more specific. Due to frequently dense connections between high first-order influence users, these users also have very high second-order influence, usually the highest in the network. For this study, we wanted to locate users with high second-order and low first-order influence. This distinction is not as simple as eliminating users with high first-order influence from the result set, as we will see in the discussion below. These users, we theorized, would include on-the-scene local reports. These might be as diverse as trusted family members or news reporters reporting on natural disasters, to the man from Abbottabad who tweeted about helicopter noise, indicating the raid that successfully targeted Osama Bin Laden. Identifying and tracking these users would remove the delay associated with them being retweeted by high-profile news sources and would remove any filtering being done by those sources. To locate these users, we developed and tested a suite of potential metrics.

The first two metrics we tested tried different approaches to isolating these users. Our first metric, Decremented Second-Order Centrality (see Equation 1), is a straightforward naïve approach. We computed a simple approximation of second-order influence for each user by summing the in-degree of all users that mentioned the user in question (in the future we will refer to this value as the unweighted second-order degree centrality), then subtracted the degree of the user. This proved insufficient, for several reasons. First, the natural degree distribution of the network was broad enough that the average degree of the network was significantly greater than 1 (we considered subtracting the degree of the central user multiplied by a coefficient equal to the average degree of the network, but rejected this as too closely associated with a single data set, making it harder to generalize). Second, we found that users with high first-order influence often had considerably higher second-order influence than the users we were looking for, leading to them continuing to outperform our target users even in this metric. As a result, this metric continued to identify first-order central individuals, as many classic metrics do.

$$\left(\sum_{user \in adj(x)} \deg(user)\right) - \deg(x)$$

Equation 1: Decremented Second-Order Centrality

Our second metric, Scaled Second-Order Centrality (see Equation 2), took a different tack. Rather than the linear difference of second-order and first-order degree centralities, we took the ratio of the two (or equivalently, we calculated the average degree of users who referred to the central user). This metric proved much more successful. The comparatively larger second-order influence of high-profile users was reduced, effectively, to simply the number of high-profile users referred to them, which was considerably lower than the values for low-profile users referred to by high-profile ones. This metric provided good progress towards our goals, but had several drawbacks, most notably that it did not provide higher values for users that were mentioned repeatedly by high-profile users or for users that were mentioned by multiple high-profile users. This metric may be useful in identifying a broad base of information sources or collaborators of first-order influencers, but does not contribute insight into ranking or categorizing their contributions.

$$\sum_{user \in adj(x)} \deg(user) \deg(x)$$

Equation 2: Scaled Second-Order Centrality

Based on the results of these two metrics, we iterated by developing a new pair of metrics, each a refined version of one of the previous two. We began with a refined version of the Decremented Second-Order Centrality metric, which we called Logarithm-Based Second-Order Centrality, illustrated in Equation 3. Rather than effectively subtracting 1 from every adjacent user's degree before adding it to a running sum, we instead took the logarithm of each user's centrality before adding it to the sum. This still had the effect of removing all nodes with degree 1 from the sum, but also reduced the impact of the medium-profile users. Ultimately, however, even this proved ineffective, as it failed to resolve the problems that stemmed from the comparatively higher second-order influence of high-profile users over the users for which we were looking. This metric may be useful, however, in enhancing the contrast between high-influence and low-influence individuals in cases where a distinction does not need to be made between first- and second-order influence.

$$\sum_{user \in adj(x)} \log(\deg(user))$$

Equation 3: Logarithm-Based Second-Order Centrality

Our refined version of the Scaled Second-Order Centrality metric was called Edge-Weighted Second-Order Centrality, illustrated in Equation 4. To calculate it, we added new information to the graph by giving edges (based on mentions) a weight equal to the log of the number of mentions between the relevant pair of users. From this, we calculated a weighted degree average by multiplying each adjacent user's degree by the weight of the edge between them and the central user before adding them to the sum (we call this sum, before it is divided by the central node's degree, the weighted second-order degree centrality). This successfully promoted users repeatedly mentioned by the same high-profile source over those who were only mentioned a single time. This approach still resulted in the highest scores being achieved by users only ever referenced by a single high-profile user, but this was considered acceptable, as ultimately any metric that does not suffer from this problem requires a way to categorize nodes into "low-

profile" and "high-profile," and as both iterations of the linear difference metric proved, this ends up being a highly network-specific question.

$$\frac{\sum_{user \in adj(x)} \deg(user) \cdot edgeweight(user, x)}{\deg(x)}$$

Equation 4: Edge-Weighted Second-Order Centrality

After comparison of all four examined metrics, we determined that Edge-Weighted Second-Order Centrality generally was most successful at identifying users with high second-order influence and low first-order influence. These metrics are tested in the following section.

4. Metric Results

We applied the trust metrics to the anonymized OWS data set, which contained 671MB worth of tweets (approximately two million tweets), and compared the metrics against one another to test their ability to identify promising second-order influencers. This data set was fully anonymized and provided by the study's sponsor at the Office of Naval Research. The running time to analyze the data set was under five minutes. Each of the algorithms rely upon a summation of the degree of surrounding users, meaning that running an algorithm on a network of n nodes will require an analysis of each node's degree, and for each node, those degrees must be operated on and summed. This process scales linearly, meaning that these metrics are likely to prove effective for all conceivable applications.

We present the results here in four tables with each corresponding to the four algorithms developed: (1) Decremented Second-Order (DSO) Centrality, (2) Scaled Second-Order (SSO) Centrality, (3) Log-Based Second-Order (LBSO) Centrality, and (4) Edge-Weighted Second-Order (EWSO) Centrality. Each table focuses on a separate metric, with the highest value nodes from that algorithm presented in order. The node IDs and values for the other algorithms are provided as well to allow for comparison of results across algorithms and to identify correlations among their results.

Table 1 demonstrates the results from the naïve metric, Decremented Second-Order Centrality. The table lists the nodes with the highest degree, decremented by one (the highest-degree node in the OWS data set was node 19 with 561 links). Little insight is provided by the first table alone, but having this metric provides insight for the subsequent analyses.

Table 1: Trust metrics ordered by Decremented Second-Order (DSO) Centrality

Node ID	**DSO**	SSO	LBSO	EWSO
19	560	1.43	12.2	2.77
1818	391	1.18	5.26	1.74
62	386	2.018	10.9	3.32
367	236	3.19	12.0	5.86
2035	207	0.207	0.0414	0.207
259	195	1.18	2.50	1.56
20	187	3.39	6.80	4.23
826	181	2.58	3.96	3.39
212	171	2.06	3.90	3.29
1194	163	3.57	9.74	6.41

The top results for Scaled Second-Order Centrality are illustrated in

Table 2. The top node, 14043, also correlates with high values for Log-Based and Edge-Weighted Second-Order Centrality. Of note, all of the top Scaled nodes have a value of zero for their Decremented cases. Reviewing this conclusion, we see that the Scaled Second-Order Centrality algorithm is working as designed: it focuses its attention on individuals that do not have high degree, but nonetheless have high influence. However, as noted in Section **Error! Reference source not found.**, this algorithm does not provide higher values for users that are mentioned repeatedly by high-profile users or for users that are mentioned by multiple high-profile users. This point will not become fully clear until we study Edge-Weighted Second-Order Centrality.

Table 2: Trust metrics ordered by Scaled Second-Order (SSO) Centrality

Node ID	DSO	**SSO**	LBSO	EWSO
14043	0	188	2.68	394
14006	0	188	1.28	188
5316	0	145	1.16	145
16907	0	123	1.09	123
14627	0	123	1.09	123
14314	0	123	1.09	123
14088	0	123	1.09	123
14087	0	123	1.09	123
13525	0	123	1.09	123
12827	0	123	1.09	123

Table 3 provides an overview of results ordered by Log-Based Second-Order Centrality. The intent of this algorithm was to "filter out" or lessen the impact of linkages to other nodes with low first-order influence by summing the logarithms of neighbors' first-order influences. This method was generally successful, but looking at the top results, many top Log-Based influencers also have high Decremented results. This result means that the Log-Based algorithm is working as advertised, but fails to perform a second step of filtering out users who are high first-order influencers in addition to second-order influencers. In practice, additional predicate logic can filter out first-order influencers to truly reveal the non-obvious trusted sources, but that approach may be subject to scalability issues.

Table 3: Trust metrics ordered by Log-Based Second-Order (LBSO) Centrality

Node ID	DSO	SSO	**LBSO**	EWSO
344	122	5.21	14.0	11.1
19	560	1.43	12.2	2.77
367	236	3.19	12.0	5.86
62	386	2.01	10.9	3.32
4452	63	11.8	10.6	17.8
385	81	6.13	10.6	15.9
1194	163	3.57	9.74	6.41
2206	144	5.10	9.71	7.41
436	133	4.09	9.20	6.26
1309	66	5.45	8.70	12.0

Table 4 provides an overview of results ranked by Edge-Weighted Second-Order Centrality. Results are similar to the Scaled results (see

Table 2), as one might expect since the two algorithms are related. Similar to the Scaled results, the highest Edge-Weighted results contain exclusively nodes with minimal first-order influence, which is a positive result. The intention of adding edge weighting in this algorithm was to promote users who were mentioned frequently by individuals with high influence. The results indicate that we succeed in promoting those users, since the table reveals that users must satisfy three general criteria simultaneously: (1) they must have low first-order influence; (2) they must have high Scaled influence (a direct but naïve indicator of second-order influence); and (3) they must have high Log-Based influence (a more sensitive metric of influence). Based on these results, we posit that Edge-Weighted Second-Order Centrality is a very viable metric to identify non-obvious sources of trustworthy and high-value information.

Table 4: Trust metrics ordered by Edge-Weighted Second-Order (EWSO) Centrality

Node ID	DSO	SSO	LBSO	**EWSO**
14043	0	188	2.68	395
11455	0	123	2.61	294
11454	0	123	1.86	208
2778	0	123	1.86	208
1062	0	123	1.86	208
14006	0	188	1.28	188
5316	0	145	1.16	145
16907	0	123	1.09	123
14627	0	123	1.09	123
14314	0	123	1.09	123

5. Conclusion

The proliferation of social media has changed the information landscape for a variety of events. In this study, we focus on a universal contributing factor to information value: trust. To assess the trustworthiness of social media data, we designed a set of metrics based on the theory of trust in information and tested those metrics on anonymized Twitter data. These metrics are likely to help analysts identify trustworthy information more quickly than manual inspection, and more generally contribute to the science of representing the behavior of humans interacting through social media. Future work may consider the application of these algorithms to various domains, including responses to natural disasters and crises that result from violent crowds. Future work may also consider comparing the contributions of the automated metrics to trustworthiness or information value metrics generated by human experts, and refining the automated metrics by including various aspects of expert judgments. As mentioned above, these metrics are not intended to provide a single score for the trustworthiness of given tweets; rather, these metrics provide one *component* of the notion of trust, which is multi-faceted. Additional research may seek to quantify other aspects of trust, providing analysts with the ability to quickly

synthesize more holistic perspectives on trustworthiness in social media.

References

Adali, S., R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B. K. Szymanski, W. A. Wallace and G. T. Williams. 2010. "Measuring Behavioral Trust in Social Networks." IEEE International Conference on Intelligence and Security Informatics (ISI).

Arrow, K. J. 1974. *The Limits of Organization*. First Edition, New York, NY: W. W. Norton & Company.

Bouma, J., E. Bulte, and D. van Soest. 2008. "Trust and Cooperation: Social Capital and Community Resource Management." *Journal of Environmental Economics and Management* 56(2):155-166.

Coleman, J. S. 1990. *Foundations of Social Theory*. Cam bridge, MA: Harvard University Press.

Fehr, E. 2009. "On the Economics and Biology of Trust." *Journal of the European Economic Association* 7(2–3):235-266.

Fehr, E., and S. Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* 14(3):159-181.

Hahn, T., P. O lsson, C. Folke, and K. Johansson. 2006. "Trust-building, Knowledge Generation and Organizational Innovations: The Role of a Bridging Organization for Adaptive Comanagement of a Wetland Landscape around Kristianstad, Sweden." *Human Ecology* 34(4):573-592.

Henry, A.D., and T. Dietz. 2011. "Information, networks, and the complexity of trust in commons governance." *International Journal of the Commons*, 5(2).

McPherson, M., L. Sm ith-Lovin, and J. M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415-444.

Munro, G. D., and P. H. Ditto. 1997. "Biased Assim ilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information." *Personality and Social Psychology Bulletin* 23(6):636-653.

Poteete, A. R., M. A. Janssen, and E. O strom . 2010. *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*. Princeton, NJ: Princeton University Press.

Rasmusen, E. 2001. *Games and Information: An Introduction to Game Theory*. 3rd ed., Malden, MA: Blackwell Publishers.

Renn, O . 2008. *Risk Governance: Coping with Uncertainty in a Complex World*. London, UK: Earthscan.

Ruttan, L. M. 2006. "Sociocultural Heterogeneity and the Commons." *Current Anthropology* 47(5):843-853.

Sabatier, P., ed. 1999. *Theories of the Policy Process*. Boulder, CO : Westview Press.

Siegrist, M., T. C. Earle, and H. Gutscher, eds. 2007. *Trust in Cooperative Risk Management: Uncertainty and Scepticism in the Public Mind*. London, UK: Earthscan.

Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock. 2006. "New Specifications for Exponential Random Graph Models." *Sociological Methodology* 36:99-153.

Sztompka, P. 1999. *Trust: A Sociological Theory*. Cambridge, UK: Cambridge University Press.

Vollan, B. 2008. "Socio-Ecological Explanations for Crowding-Out Effects from Economic Field Experiments in Southern Africa." *Ecological Economics* 67(4):560-573.

Young, O . 2002. "Institutional Interplay: The Environmental Consequences of Cross-Scale Linkages." *The Drama of the Commons.* Washington, DC: National Academy Press.