Building Trusted Social Media Communities:
A Research Roadmap for Promoting Credible Content

Ben Shneiderman, University of Maryland--College Park (draft September 9, 2013)

Abstract: *A growing body of literature and inspirational examples provides guidance for aspiring social media community leaders. We know that design principles for websites can make a substantial difference in getting first-time users to return and to trust commercial, academic, government, and other websites. By contrast, building credible social media communities requires large numbers of regular content contributors guided by inspirational and committed leaders. This paper offers a defining framework for discussing the social, technical, and content foundations that encourage trusted contributors to contribute credible content to social media communities. Each component of the framework -- the trusted contributors, credible content, reliable resources, and responsible organizations -- can be undermined. Therefore, researchers and community leaders who attend to each component have a higher chance to produce positive outcomes. This framework provides a road map for research on and management of credible communities.*

Introduction

Trusted contributors who provide *credible content* are vital nutrients for successful social media communities. When community members can rely on responses to questions, restaurant reviews, or healthcare recommendations they may benefit personally and be more likely to help others. Social capital as well as tangible economic benefits grow when good deeds are rewarded and malicious actions are suppressed. In addition, *reliable resources* of software, hardware, servers, and networks provide the technical foundation, while *responsible organizations* ensure a robust socio-technical foundation.

Techniques for assessing credibility and design principles that encourage trustworthy behavior are still emerging as the web, mobile, and social technologies mature. Early studies of website credibility focused on surface features such as spelling errors, willingness to provide contact information, professional appearance, rapid response, recognizable domain name, recency of content, and volume of information [4, 5, 6, 7, 8, 20]. Later work began to emphasize external markers such as verifiable seals of approval (e.g. eTrust, BBB, Microsoft MVP), public reputations based on long-term performance (e.g. eBay, Amazon), references from other users (e.g. likes, confirmations, badges, karma points), and visible histories of activities (e.g. Wikipedia edits, Amazon reviews). These more complex systems are still maturing as community site managers refine designs to promote more credible content that is less subject to deceptive practices [11, 14, 15, 17, 28].

The distinctive open nature of social media communities means that millions of people may post content such as reviews, answers to questions, videos, or comments on blogs. This significant design choice opens up participation broadly, but presents new challenges to researchers and community leaders. Off-topic postings, links to commercial or pornographic sites, and libelous attacks can easily disrupt and undermine a thriving community. The volume of posting means that centralized review is difficult, so automated and social approaches to ensuring credibility are necessary.

Dangers exist from those who build reputations artificially or legitimately with the goal of ultimately providing misleading advice [15]. These botnet-facilitated deceptions and cleverly designed moles require more sophisticated filters to detect. Simpler, but effective, threats come from individuals self-promoting their work, companies surreptitiously promoting their products, or political actors undermining opponents. Criminals and terrorists may be few in number but they have more troubling agendas, and they are often well-organized and knowledgeable. The disturbing reality is that trust is fragile, so that a small fraction of misleading or malicious postings can undermine an otherwise trustworthy community.

While small social media communities are rarely attacked, as they succeed they become more attractive targets, requiring increasingly diligent monitoring to preserve credible content. Wikipedia has developed an especially rich set of protections, since small slips become newsworthy stories that can dramatically undermine a long history of positive reputation. As more people depend on social media communities for travel, health, financial, and legal information, increased research and greater diligence on the part of community leaders is necessary.

A research agenda that addresses all these threats will produce a broad range of recommendations. However, traditional controlled laboratory experiments have little relevance in the large bustling world of social media communities. Reductionist models are less relevant and the number of uncontrollable variables is large. At the same time, interventions in functioning systems can be difficult to arrange and have their own risks. Therefore, partnerships between industry system managers and academic researchers could prove to be beneficial. By combining applied and basic research, which is informed by practical and theoretical frameworks, high impact outcomes seem possible. Repeated case studies using design interventions produce data that can support theories, principles, and guidelines. Such systematic interventions in working systems may prove to be the most valuable approach. Of course, automated logging when combined with ethnographic observations, in-depth interviews, and validated surveys have the potential to produce actionable research results.

Research on scalable organizational structures and processes are a further opportunity. Just as large organizations must have a hierarchy, or other structure, online communities will need to have multiple levels of management and leadership. The Reader-to-Leader Framework suggests how multiple levels of participation can be designed into systems [21]. Successful communities have large number of readers of the content, but often the number of content contributors may be in the neighborhood of one percent of the readers. Those who become active collaborators, engaging in discussions with other contributors are a still smaller circle. Those who rise to leadership positions to guide design processes, cope with problems, and mentor novices is a still narrower circle, but an essential component to a thriving community. In large communities, such as Wikipedia, there are many formal policies and evolving norms, so there is often a great deal for newcomers or aspiring leaders to learn. Creating motivations for readers to become contributors and then collaborators, and eventually a leader is crucial. Then providing recognition for those who contribute actively or collaborate productively are further challenges. Research opportunities abound for those seeking to study how visible recognition of positive contributions (downloads, likes, retweets, etc.) and rewards for substantial efforts (leaderboard of most prolific contributors, selection as a Wikipedia Featured Article, Most Valuable Professional awards).

The leaders help set inspirational agendas, promote behavioral norms by their examples, take the community into new directions, and deal with a wide variety of threats. Successful communities must

develop leaders who create resilient social structures to deal with serious threats from hackers who maliciously violate privacy, attack servers, vandalize content, or provide misleading content

Even large communities can go astray, failing to attract, motivate, and recognize contributors adequately. These communities can also face challenges from malicious participants who wish to subvert the community for their own purposes. Worse still internal dissent, corrupt leaders, or failure to serve stakeholders can rapidly undermine trust, which may be difficult to recover. This was the scenario for Digg's failure (http://www.computerworld.com/s/article/9214796/Elgan_Why_Digg_failed). Therefore, independent oversight by external bodies with high reputation offers a proven approach for corporations, government agencies, or universities that could be valuable in social media communities.

Previous work on web credibility guidelines provides a foundation for social media community credibility, but the shift from a centralized web construction model to an open participatory community environment introduces many new concerns.

The Stanford Web Credibility Project (http://en.wikipedia.org/wiki/Stanford_Web_Credibility_Project) compiled 10 reasonable guidelines [5, 6, 7]:

1. Make it easy to verify the accuracy of the information on your site.
2. Show that there's a real organization behind your site.
3. Highlight the expertise in your organization and in the content and services you provide.
4. Show that honest and trustworthy people stand behind your site.
5. Make it easy to contact you.
6. Design your site so it looks professional (or is appropriate for your purpose).
7. Make your site easy to use—and useful.
8. Update your site's content often (at least show it's been reviewed recently).
9. Use restraint with any promotional content (e.g., ads, offers).
10. Avoid errors of all types, no matter how small they seem.

Others have extended the list of web credibility guidelines up to 39 items (http://conversionxl.com/website-credibility-checklist-factors/#), such as showing staff bios and photos, client lists, testimonials, and trust marks. A workshop devoted to web credibility contains a set of early helpful papers (http://projects.ischool.washington.edu/credibility/).

These are valuable points of departure but the open nature of social media communities presents far greater challenges for researchers, community leaders, and community members for are seeking credible content. Research on trust in social media communities [9] is a growing topic, which deserves further attention.

Framework for credible communities

Like the proverbial elephant, there are many ways to think of social media communities. Sociologists may focus on the bounded nature of community members and seek to ensure that only trusted contributors participate [30]. Natural language researchers may study the inherent sentiment or linguistic patterns in the millions of posts, looking for indicators of credible content. Privacy and

security analysts want to certify the software, control the devices, restrict access to servers, and protect their networks, while social theorists focus on responsible organizations such as professional societies, corporations, and government agencies.

There are undoubtedly more ways of thinking about credible communities, but these four components (Figure 1) already constitute a large and complex socio-technical system that provides a plethora of research opportunities. At the same time, this four-component framework gives community leaders and members a way to organize their discussions and actions so as to raise their credibility. Each component suggests research tasks, the need for operational tools, and the development of guidelines for community leaders and members. For management effectiveness quality metrics will be needed to monitor changes and assess the impact of systematic interventions.

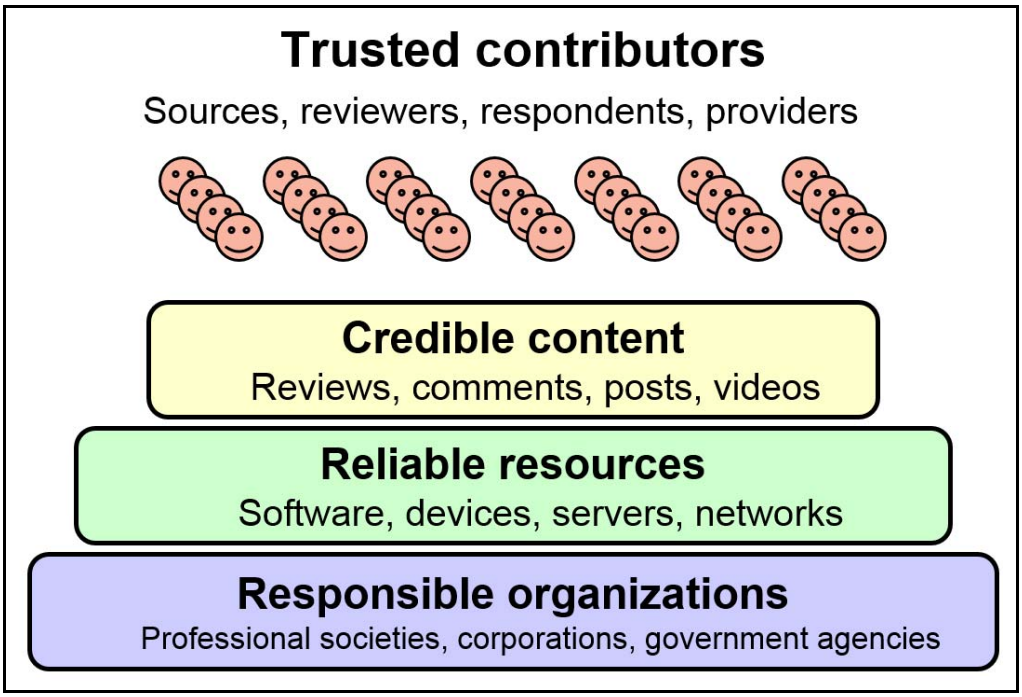

Figure 1: A framework for analysis of social media communities. Ideally, trusted contributors provide credible content, that is delivered by reliable resources, guided by responsible organizations. However contributors may be mis-informed, biased, or malicious, so their content is not credible. Similarly, physical resources can be undermined and organizations may be subverted or become corrupt.

A few initial thoughts may trigger deeper thinking and constructive work on (1) trusted contributors, (2) credible content, (3) reliable resources, and (4) responsible organizations.

(1) Trusted Contributors

Every community would like to have only trusted contributors, but the rough reality is that many contributors are mis-informed even if they are well-intentioned. They can give misleading medical advice or incomplete financial information, which could have devastating effects. Second, contributors

may be biased so they present only favorable book reviews or report only good restaurant experiences. Third, contributors may be maliciously seeking to undermine a competitor's products or a political opponent's reputation.

Many strategies are being tried to ensure that only trusted contributors participate, such as raising the barriers to entry for contributors by requiring a log-in (no anonymous contributions), identity verification, background check, probation periods, and public performance histories. Greater transparency about who the contributors are and what their past is has the potential to increase trust in their future contributions.

Ancient social processes are finding new instantiations in online communities to help ensure trusted contributors. Some communities require recommendations from members to admit new members, a waiting period before contributions are accepted, or several stages of membership so that novices have limited privileges, which are increased as positive contributions are made. However, research to validate, measure, and refine these techniques will be necessary to support practice and develop effective social theories.

Network analysis to reveal past histories of troubling relationships with known malefactors could be a powerful approach [9, 10, 11, 30]. In some cases, such as with Twitter, follower and following relationships are accessible so deeper understanding of social relationship is possible, but clever users have developed strategies to appear trustworthy or cover troubling histories. Research on advanced network analysis techniques could improve their efficacy and resistance to subversion [23]. Trustworthy contributors are likely to be related to other trustworthy contributors, but developing a metric based on networks would be a helpful strategy.

(2) Credible Content

The core of community credibility is credible content: movie reviews, responses to technical questions, blog posts about travel destinations, how-to videos, and much more. Verifying that each content offering is credible is an enormous and impossible task, especially as the volume and pace grows. Even within the range of credible content there is a wide range in quality [12, 19] of content, ranging from brief notes to detailed commentaries with evidence to support claims. Studies of question answering websites have shown that those websites that require question askers to pay for answers produce higher quality answers.

While encouraging high quality is one research goal, another is filtering out off-topic, inappropriate, or unhelpful postings. Spam filters for email have been refined enough to work quite reliably and rapidly, but that experience is only partially applicable to building credible communities. Tracking contributors and comparing content against blacklist databases of names and spam messages are basic approaches, which could be adopted for social media communities [22]. In addition, research on sophisticated text analysis of individual content items and comparisons with similar items can all help to ensure that only credible content is ever made public. However, these filters are imperfect and attackers will become increasingly sophisticated [3, 14]. Therefore follow-up verifications and retrospective analyses of all content submitted by a contributor can be helpful.

Social processes such as community confirmation by votes or likes and mechanisms for community members to challenge content can also be beneficial. These processes all build awareness of the threats

and a greater devotion to building a credible community. Here again, anecdotal evidence is encouraging, but systematic research and innovative interventions will be helpful. For example, changing from simple "Likes" to allow "Respect" could allow community members to make more nuanced comments on political content [27]. Community managers who wish to ensure credible content face additional challenges in dealing with political discussions or debates over controversial subjects such as climate change or abortion. Content may be seen as credible by some readers, but not by others, often leading to hostile debates that cannot be easily resolved.

While social media community designers are increasingly adding features to promote credible content, there are also leadership strategies to motivate community members to participate in credibility-supportive ways, while discouraging malicious actors. Inspirational leaders who express visionary beliefs about their community can encourage members to be more active in ensuring credible content. These leaders can promote social norms by their examples or praising actions of members, possibly tied to motivations such as altruism, egoism, collectivism (commitment to helping a community), and principlism (devotion to doing good deeds) [2, 29]. They can also arrange social processes by which the members adopt and enforce policies about content, with punishments for violators, and dispute resolution processes to deal with naturally emerging differences. A well-managed community with devoted members who care about their community may be able to inoculate itself against threats and show resilience after attacks or damaging episodes. Research that tracks threats, attacks, and resilient responses could provide valuable guidelines for managers and predictive theories.

(3) Reliable Resources

A credible community depends on reliable resources, including trustworthy software, dependable devices, well-managed servers, and secure networks. Each of these software and hardware components has large research communities devoted to self-improvement, but since all these components are needed to produce a credible community, there are many paths to failures. Bug-free software, secure devices, non-stop servers, and private networks are all fantasies promoted by many well-intentioned people, but the reality of these complex systems is that they are dangerously vulnerable [13, 14].

Strong privacy protection builds trust and credibility. Users who fear that their identity, personal data, address, or photo will be exposed beyond the range of those who they grant permission will resist participating or provide only partial information. Research on privacy is a vast topic already, with progress being made about enabling users to understand and specify their privacy requirements [1].

The realistic response is to strive for reliable resources, while continuously monitoring performance and repairing problems promptly. Another part of a realistic response is to make honest statements to all stakeholders about the vulnerabilities, report openly about failures, and invite efforts to make improvements. Active research continues on these issues because so much of every country's national infrastructure depends on reliable resources. Social media communities have some special needs because of the large and rapidly growing numbers of users, the high variance between normal and peak usages, and because malicious actors often target these resources.

(4) Responsible Organizations

We all like to believe that our large international, national, or local organizations are responsible, accountable, and even liable for failures. We all like to believe that these organizations are run by

informed leaders acting on behalf of their members with integrity and honesty. Once again the reality falls far behind the expectations, producing organizations that are corrupt, self-serving, or incompetent.

While there are no guaranteed methods to ensure responsible organizations, the goal is an important one that needs discussion and research. Internal audits, transparent processes, and open reporting of performance are good starts. However, independent oversight by trusted external organizations is still a valuable approach. Better Business Bureau Online, eTrust.org, and truste.com offer some approaches that could help build more credible communities, but research on still newer approaches will be beneficial.

Independent oversight can occur in many ways. Continuous oversight by trusted individuals or organizations is effective but expensive. A less costly approach, annual reviews, such as corporate audits, are commonly done, but vary in their effectiveness. Strong annual reviews by informed panels who have open access to historical records can lead to valuable reports and recommendations, but the follow-up to ensure that recommendations are followed is vital. Finally, review panels when disasters occur, such as in airline crashes, can lead to recommendations to reduce future threats, but only if conducted in an open environment with full disclosure of reports [18].

Conclusion

The promise of social media communities is that they lower barriers to participation so as to create valuable resources, give assistance where needed, and promote more informed decisions among billions of users. However, the reality is more troubling. Mis-informed, biased, and malicious contributors could produce harmful content that would undermine trust enough to destroy the value of these communities. Other threats such as corrupt leaders and internal strife can also undermine otherwise credible communities.

A substantial research effort will be needed to raise the possibility that outcomes will be positive. The research agenda offers rich possibilities for many disciplines and inter-disciplines. Multiple research methods, including novel ones, will be needed because of the tightly-interrelated nature of social media communities, which defy reductionist approaches. Carefully monitored interventions and rigorous case studies are likely to be more valuable than controlled experiments. Furthermore, research projects that combine basic and applied goals, practical and theoretical approaches, and mission-driven and curiosity-driven aspirations seem more promising than fragmentary efforts [24].

At the same time, designers of social media communities will have to work diligently to produce effective user interfaces, supported by reliable resources, so that community leaders and members can contribute credible content while they help raise the quality of everyone's contributions. There is also research to be done by software, hardware, and network designers, as well as by organizational designers. Responsible organizations can have powerful impacts, especially when their actions encourage every individual contributor to produce credible content.

Acknowledgements

Thanks to the conference organizers, Sorin Matei and Elisa Bertino, for inviting this keynote and encouraging my work on credibility. I appreciate helpful comments from Brian Butler, BJ Fogg, Jennifer Golbeck, Michael Hicks, Itai Himelboim, Jonathan Lazar, Alan Mislove, and Jennifer Preece.

References

1. Baden, R., Bender, A., Spring, N., Bhattacharjee, B., and Starin, D., Persona: an online social network with user-defined privacy, *ACM SIGCOMM Computer Communications Review 39,* 4 (Aug. 2009), 135-146.
2. Batson, C.D., Ahmad, N., and Tsang, J., Four motives for community involvement, *Journal of Social Issues* (58) 3 (2002), 429-445.
3. Beutel, A., Xu, W., Guruswami, V., Palow, C., and Faloutsos, C., CopyCatch: Stopping group attacks by spotting lockstep behavior in social networks, *Proc. 22nd International World Wide Web Conference (WWW'13),* Rio de Janeiro, Brazil (May 2013).
4. Ceaparu, I., Demner, D., Hung, E., Zhao, H., and Shneiderman, B., "In Web We Trust": Establishing strategic trust among online customers, In Rust, R. and Kannan, P. K. (Eds), *E-Service*, M. E. Sharpe Pubs, Armonk, NY (2002), 90-107.
5. Fogg, B.J., *Persuasive Technology: Using Computers to Change What We Think and Do,* San Francisco, CA: Morgan Kaufmann (2002).
6. Fogg, B. J., Marshall, J., Osipovich, A., Varma, C, Laraki, O., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., and Treinen, M., Elements that affect web credibility: early results from a self-report study, *Proc. CHI '00 Extended Abstracts on Human Factors in Computing Systems*, [doi>10.1145/633292.633460]
7. Fogg, B. J. and Tseng, H., The elements of computer credibility, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, (1999), 80-87.
8. Friedman, B., Kahn, P.H. Jr., and Howe, D. C., Trust online, *Communications of the ACM.43*, 12 (Dec 2000), 34-40. [doi>10.1145/355112.355120]
9. Golbeck, J., *Computing with Trust*, Springer, Berlin (2009).
10. Golbeck, J., *Analyzing the Social Web*, Morgan Kaufmann (2013).
11. Hansen, D., Shneiderman, B., and Smith, M., *Analyzing Social Media Networks with NodeXL*, Morgan Kaufmann (2010).
12. Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A., Predictors of answer quality in online Q&A sites. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, ACM (2008), 865-874).
13. Jim, T., Swamy, N., and Hicks, M., Defeating scripting attacks with browser-enforced embedded policies, *Proc. International World Wide Web Conference (WWW)* (May 2007), 601-610.
14. Kakhki, A. M., Kliman-Silver, C., and Mislove, A., Iolaus: Securing online content rating systems, *Proc. 22nd International World Wide Web Conference (WWW'13),* Rio de Janeiro, Brazil (May 2013).
15. Kittur, A., Suh, B., and Chi, E. H., Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia, *Proc. ACM 2008 conference on Computer supported cooperative work* (2008).
16. Lazar, J., Meiselwitz, G., and Feng, J., Understanding web credibility: A synthesis of the research literature, *Foundations and Trends in HCI 1*, 2 (2007), 139-202.

17. McKnight, D. H. and Kacmar, C., Factors and effects of information credibility, *Proc. 9th International Conference on Electronic Commerce*, August 19-22, 2007, Minneapolis, MN, USA
18. National Academies Committee on Technical and Privacy Dimensions of Information for Terrorism Prevention and Other National Goals, *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*, National Academies Press, Washington, DC (2008), Available at: http://www.nap.edu/catalog.php?record_id=12452
19. Nichols, J., Zhou, M., Yang, H., Kang, J-H., and Sun, X. H., Analyzing the quality of information solicited from targeted strangers on social media, Proc. 2013 Conference on Computer Supported Cooperative Work (2013), 967-976.
20. Pirolli, P., Wollny, E., and Suh, B., So you know you're getting the best possible information: a tool that increases Wikipedia credibility, *ACM CHI* (2009).
21. Preece, J. and Shneiderman, B., The Reader-to-Leader Framework: Motivating technology-mediated social participation, *AIS Transactions on Human-Computer Interaction 1*, 1 (2009), 13-32, http://aisel.aisnet.org/thci/vol1/iss1/5/
22. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Menczer, F., Detecting and tracking political abuse in social media, *Int'l Conf on Web Search and Social Media (ICWSM)*, IEEE (2011).
23. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F., Truthy: mapping the spread of astroturf in microblog streams, *Proc. Int'l World Wide Web Conference (Companion Volume)* (2011), 249-252.
24. Rubin, V. L., On deception and deception detection: content analysis of computer-mediated stated beliefs, *Proc. 73rd ASIS&T Annual Meeting* (2010).
25. Shneiderman, B., Designing trust into online experiences, *Comm. ACM*, 43, 12 (2000), 57-59.
26. Shneiderman, B., Toward an ecological model of research and development, *The Atlantic* (April 2013). http://www.theatlantic.com/technology/archive/2013/04/toward-an-ecological-model-of-research-and-development/275187/
27. Stroud, N. J., Muddiman, A., and Scacco, J., Framing comments in social media, National Communication Association, Political Communication Division, Washington DC (November 2013).
28. Vega, L. C., Sun, Y.-T., McCrickard, D. S., and Harrison, S., Time: a method of detecting the dynamic variances of trust, *Proc. 4th Workshop on Information Credibility* (2010).
29. Violi, N., Shneiderman, B., Hanson, A., and Rey, P., Motivation for participation in online neighborhood watch communities: An empirical study involving invitation letters, *Proc. IEEE Conference on Social Computing*, IEEE Press, Piscataway, NJ (October 2011).
30. Westerman, D., Spence, P. R., and Van Der Heide, B., A social network as information: The effect of system generated reports of connectedness on credibility on Twitter, *Computers in Human Behavior 28*, 1 (2012), 199-206.
31. Wu, J-J. and Tsang, A.S.L., Factors affecting members' trust belief and behaviour intention in virtual communities, *Behaviour and Information Technology 27*, 2 (2008), 115-125.