Modeling Online Creative Collaborations

A study of the online music writing community FAWM.ORG reveals that people who collaborate share less in common than you might think.

By Steven Dow and Burr Settles

DOI: 10.1145/2460436.2460445

What is the genesis of a great creative collaboration? From music to movies to multimedia, many creative projects involve two or more partners who team up to work their collective magic. However, we know very little about the special chemistry that spurs these artists to join forces. Our work applies a novel path-based regression technique to mine social interactions and to identify collaboration patterns in an online music writing community.

This novel application of computer science techniques leads us to a deeper understanding of creative collaboration in online communities, and can provide insights for other researchers exploring the intersection of social science and big data.

THE EMERGENCE OF ONLINE CREATIVE COLLABS

Web technologies and low-cost production tools have led to a surge in "peer production," from epic cultural efforts such as Wikipedia, to open-source software, to smaller communities that nourish more specific artistic pursuits. For example, millions of registered users have created hundreds of thousands of animated movies and online games at Newgrounds.com (an online community of animators, and one of the most heavily-trafficked sites on the Internet) [1]. Likewise, nearly three million interactive media projects have been created and shared in the Scratch online community [2].

Research shows that working with others to achieve shared goals can promote social, motivational, and emotional benefits. For example, collaboration in classroom settings can improve peer relationships, increase self-esteem, and develop perspective-taking skills [3]. Recently, we have also studied collaboration online, where we can investigate how teamwork, individual and group goals, and communication affect one another. For example, in a previous study about the online music community called February Album Writing Month (FAWM), we found newcomers who engage in one-on-one "collabs" during their first year are not only more successful at reaching their own personal songwriting goals, but also go on to behave in more community-favorable ways, including commenting on others' music or donating money to the site [4]. This suggests collaborative efforts can improve outcomes for both individuals and the community as a whole. Given the benefits, we sought to understand the social factors that affect the formation of online creative collabs, and to eventually develop new tools, technologies, and best practices to help these communities flourish.

> **Our goal in the current project was to build a model that can help explain and predict how collaborations form in online creative communities like FAWM.**

FEBRUARY ALBUM WRITING MONTH (FAWM)

Our work focuses on an online music community called February Album

Writing Month, an annual online music event for professional, semi-professional, and amateur songwriters (http://FAWM.ORG). The community tagline is "14 Songs in 28 Days," and it revolves around a challenge to compose at least 14 songs (roughly an album's worth) during the shortest month of the year. Over the past decade, more than 7,000 participants have participated worldwide, collectively penning about 60,000 original works of music and thousands of collaborations.

The main features of the site include user profile pages, an open discussion forum, and a list of publicly posted songs where participants can find, listen to, and comment on one another's music. User profiles include short bios, links to completed songs, and a "soundboard" where others can post direct messages. The "bulletin board" style forum contains thousands of topics on music recording, sources of inspiration, regional discussions, and "collaboration classifieds" where songwriters looking to collaborate can propose projects and team up. Song pages include author-provided descriptive tags, which are mainly used to categorize songs by genre or instrument (e.g., "punk-rock" or "piano"), and community members can leave feedback in the comments section at the bottom of each song page. Songs can also be searched and browsed; for example, a few collaborations from the FAWM 2013 event are shown in Figure 1.

Figure 1. A Screenshot of the FAWM.ORG site from 2013.

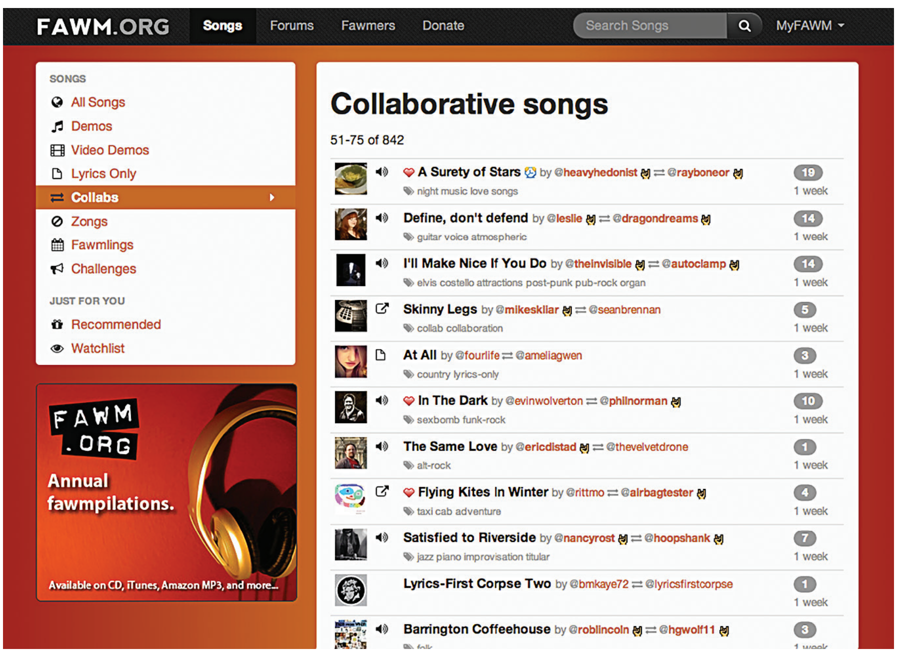
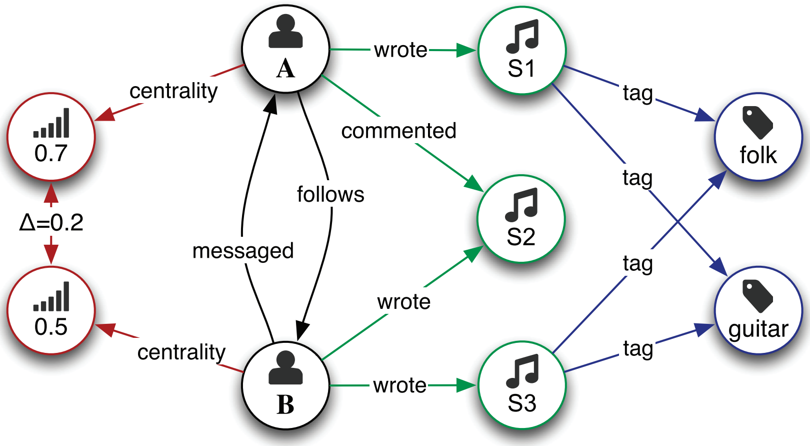


Figure 2: An illustrative example of the February Album Writing Month (FAWM) social network graph.

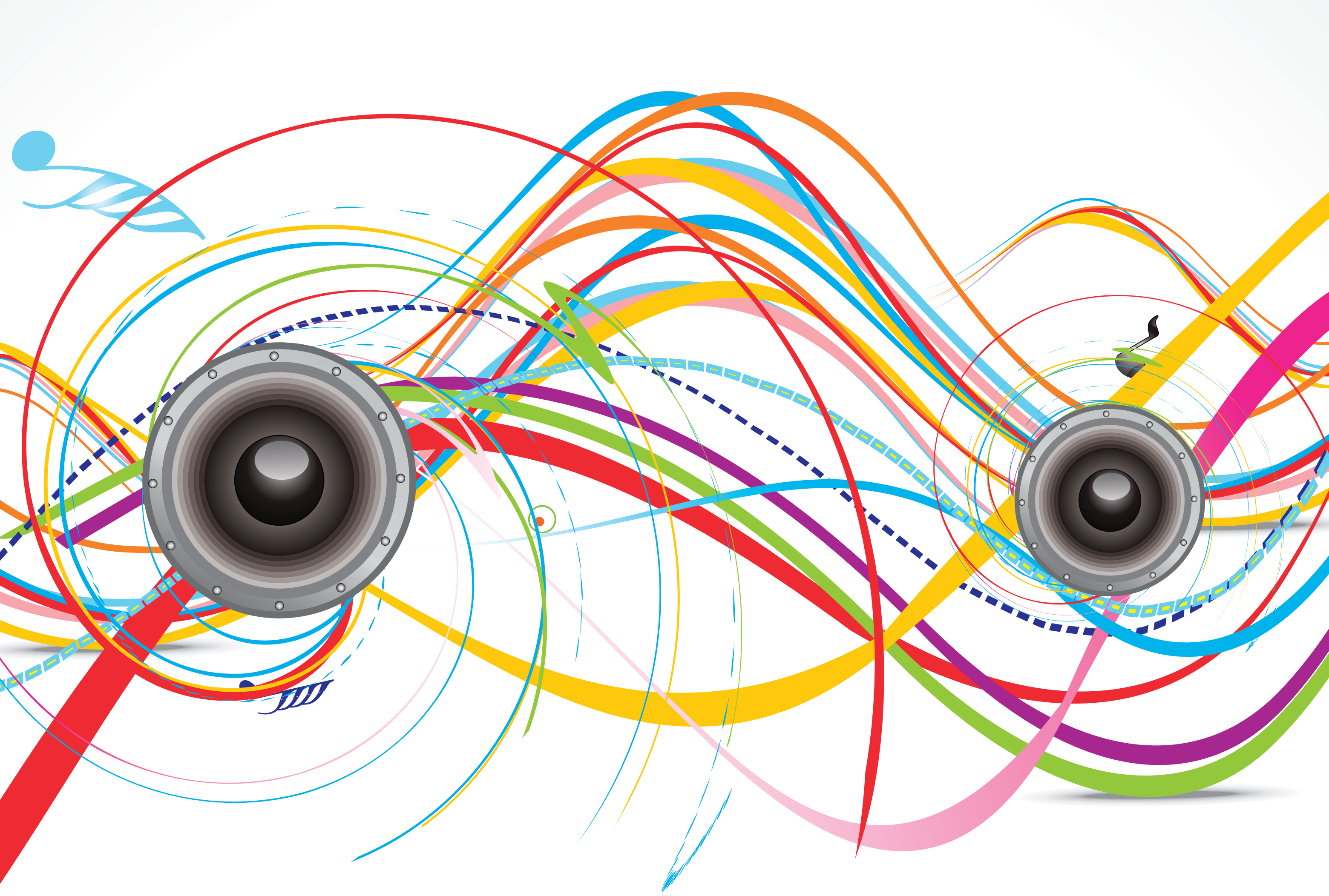


Collaborative projects in the FAWM community date back to at least 2006, when three so-called "fawmers" joined forces to each compose 14 songs about different U.S. presidents (covering all 42 presidencies in history up to that point). The collection, titled "Of Great and Mortal Men," was later released as a critically acclaimed triple-album project during the 2008 election season. The trio then toured and performed at the esteemed South by Southwest (SXSW) Music Festival for several years. This parallel, distributed-labor model of collaboration is reminiscent of open-source software and Newgrounds animation projects.

Smaller two-person collaborations became popular during FAWM 2008. Since it was a leap year, the organizers (including one of the authors here) jokingly upped the ante to "14½ Songs in 29 Days." Participants were encouraged to co-write an extra half-song. This resulted in 252 documented collaborations, or 4.4 percent of the total musical output that year. The popularity of these pairwise collaborations have grown, comprising 7.8 percent of all songs posted to the website since FAWM 2009. A notable example is "Walkthrough," by fawmers @errol and @pifie. The song's lyrics, set to ambient alternative rock music, outline the steps required to win the classic text-based computer game "Zork." The song went viral on the Internet and enjoys a certain level of notoriety among classic computer game enthusiasts.

A STATISTICAL MODEL OF CREATIVE COLLABORATION

Our goal in the current project was to build a model that can help explain and predict how collaborations form



in online creative communities like FAWM. To do that, we used a new machine learning technique to predict whether or not a "collaboration edge" will form between two users in the social network, as a function of the other ways in which they are already connected. As our data set, we used FAWM website server logs between 2009 and 2012, which included 6,116 active users, 39,103 songs posted to the site, and 3,047 documented collaborations.

The particular model we use is a new twist on logistic regression (a common statistical machine learning technique) that we call "path-based regression" [5]. Each data instance represents a pair of users *A* and *B*, the predicted output variable is whether or not the pair co-posted a collaborative song to the website (e.g., the songs in Figure 1), and the input variables are different kinds of paths that connect users through the existing social network graph.

Figure 2 shows an example of a small social network to help illustrate the idea. Suppose the model is trying to predict whether user *A* will collaborate with user *B*. One way of connecting their nodes in the social graph is through the path *A* –follows→ *B*, which means that *A* has "subscribed" to *B*'s song feed (indicating that she is interested in his work). Another path is *A* ←messaged– *B*, meaning she received a direct message from *B* on her profile. There are also longer paths, such as *A* –commented→ ♬ ←wrote– *B*, which means that *A* commented on one of *B*'s songs; or even more complex paths such as *A* –wrote→ ♬ –tag→ 🏷 ←tag– ♬ ←wrote– *B*, which means that both users have written songs tagged with a shared term (a good indication that they have shared

Figure 3: Evaluation curves comparing our path-based regression, the Adamic & Adar heuristic, and matrix factorization.

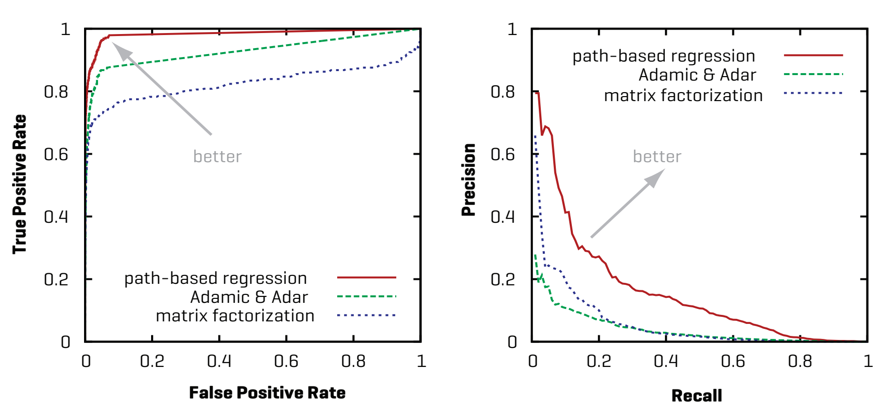


Table1: Sample weights from the path-based logistic regression predicting collab formation.

Path Variable	Weight
A ←follows– *B*	8.433
A –follows→ *B*	7.926
A ←messaged– *B*	4.935
A –messaged→ *B*	4.183
A –wrote→♫ ←commented– *B*	4.160
A –commented→♫ ←wrote– *B*	3.879
A –wrote→♫ –tag→🏷 ←tag–♫ ←commented– *B*	0.868
A –commented→♫ –tag→🏷 ←tag–♫ ←wrote– *B*	0.504
A –wrote→♫ –tag→🏷←tag–♫ ←wrote– *B*	−0.388
A –centrality→📶←Δ=0.2→📶←centrality– *B*	0.818
A –centrality→📶←Δ=0.6→📶←centrality– *B*	0.614
A –centrality→📶←Δ=0.7→📶←centrality– *B*	-0.002
A –centrality→📶←Δ=0.9→📶←centrality– *B*	-0.332
A –centrality→📶←Δ=0.8→📶←centrality– *B*	-0.455

interests in musical genre and style, or use similar instruments).

The FAWM network includes nodes derived from tables in the FAWM database such as users, songs, tags, forum topics, and the various kinds of links between them in the server logs. We were also interested in how collab formation might be affected by status within the user community, so we computed each user's eigenvector centrality—a measure of social influence similar to Google's PageRank—by using the network of communication edges. In Figure 2, for example, *A* has a centrality score of 0.7 while *B* has a score of 0.5. For each pair of centrality nodes, we added an edge representing the difference between them (e.g., $\Delta = 0.2$).

Our path-based regression method treats each type of path through the network as an input variable, whose value reflects the "strength" of that path in connecting the two users. For example, the shared-tag path has two occurrences in Figure 2: One through the "folk" tag node, and another through "guitar." The model should recognize that this path type connects *A* and *B* more strongly than it would for two other users who share only one common tag. While it is theoretically possible to tabulate these frequencies for all possible paths connecting all possible pairs of users, that would be expensive in practice and would not scale well to any large social network. Instead, we take a sampling approach based on "random walks," a very common and useful method for modern large-scale network analysis. The algorithm begins at user node *A*, selects an edge at random to arrive at a new node, and repeats until reaching user node *B* for paths up to length four. This process repeats for a finite amount of time, and the cumulative path statistics are normalized and used as inputs to describe the user pair *A* and *B*.

We gathered statistics for thousands of path types that connected many thousands of user pairs, and trained a logistic regression model to distinguish between user pairs that collaborated and pairs that did not. The path-based regression approach performs much better than other standard link-prediction methods, like the Adamic & Adar heuristic [6] and SVD-based matrix factorization [7]. Figure 3 shows ROC and precision-recall curves evaluating the predictive power of each algorithm when trained on logs from the first half of FAWM 2012, and predicting collaborations that might occur during the second half of the month.

By applying "big data" analysis techniques to social network data in online creative communities, we can gain insights about how they work and what can make them better.

FACTORS THAT AFFECT COLLAB FORMATION IN FAWM

In addition to better predictions, our path-based regression model is very interpretable compared to the other approaches. This allows us to inspect model weights associated with the different path types to examine how they might influence collab formation. Table 1 presents a few of the weights induced from the FAWM 2012 network (for other years, results are qualitatively the same in terms of sign and magnitude). To corroborate our model's predictions, we also surveyed members of the FAWM community with open-ended questions about how their collaborations formed. Our analyses of these quantitative and qualitative results reveal three main findings:

1. Communication exchanges predict collaboration. As theory suggests [8], the top six predictors of collab formation have to do with communication exchanges: following a partner's song feed, direct messaging, and commenting on a partner's songs. Survey participants confirm the importance of having a rapport with your collaborator: "The other person and I had both made comments like 'ohh we should totally do something together' … ."

2. Collabs form out of shared interests but different skills. Recall that tags are mainly used to categorize songs by genre or instrument, so paths that flow through tag nodes can be thought of as expressing a shared interest in musical style. The path *A* –wrote→ ♫ –tag→ 🏷 ←tag– ♫ ←commented– *B*, for example, means that *A* writes songs tagged with terms that are also used for the songs that *B* often comments about. In other words, *A*'s typical genre is something of shared interest to *B* according to his commenting behavior. As social identity theory implies [9], this path is a positive predictor of collaboration (as with *A* commenting on *B*'s genre). However, the path *A* –wrote→ ♬ –tag→ 🏷 ←tag– ♬ ←wrote– *B*, which means that the users often write songs in the exact same genre or style, turns out to be a negative predictor.

While this may seem contradictory or counter-intuitive at first, it suggests a more nuanced form of homophily than typically discussed in collaboration research. In particular, we speculate this reflects an exchange of skills and expertise that one party has, but the other does not. Consider this post from the FAWM 2010 discussion forums: "Sometimes I wish I had one of those screamer voices…I could do a raspy acid rage-filled rocker song. Maybe one of you rockers will take me under your AX and help me bring out the inner artistic angst??"

This member wants to stretch herself with a musical style in which she is interested but inexperienced; after looking for help in the forums, a collaborator volunteered his expertise in heavy metal. Survey respondents confirm many heterogeneous collaborations begin this way: "The collaborator, knowing my style, pitched an idea to me that I liked. We passed ideas back and forth each doing aspects [we] could do best."

Similar dynamics can manifest in other online creative communities. At Newgrounds.com, for example, collabs often form around animation projects of shared interest, but for which one partner has a production skill that the other does not (such as illustration or programming).

3. Collabs are associated with small status differences. Social network centrality also seems to play a nuanced role in collab formation. Theory predicts people are more likely to work together if they are at the same status level, and less likely if further apart [10]. Our model confirms that very different centrality scores among participants are negatively associated with collaboration (e.g., $\Delta \in [0.7–0.9]$). However, the path stating that partners

We used a new machine learning technique to predict whether or not a "collaboration edge" will form between two users in the social network, as a function of the other ways in which they are already connected.

have the exact same centrality score, *A* –centrality → 📶 ←centrality– *B*, was given a weight of zero by the model, effectively declaring it insignificant. Curiously, a difference of $\Delta = 0.2$ is the strongest positive predictor of collaboration with regard to centrality, suggesting that many partnerships form around small differences in social status.

This result is somewhat puzzling, but consistent from year to year. Survey responses provide some explanation: Users of lower rank take the opportunity to reach out to their heroes and other influential members of the community, in hopes of working together. As one fawmer put it: "I've had a FAWM crush on [her] for ages, and I was noodling on guitar and came up with something that I thought would sound awesome with her voice."

Alternatively, members of higher status sometimes reach out to less experienced songwriters or struggling newcomers, conveying a more active mentor relationship.

IMPROVING ONLINE CREATIVE COMMUNITIES

By applying "big data" analysis techniques to social network data in online creative communities, we can gain insights about how they work and what can make them better. As one exciting example, we might integrate such models of collab formation directly into the socio-technical software that support online communities: These models can make intelligent recommendations for members in search of good collaborators. Given how many FAWM members collaborate every year, and how many of them are willing to be matched up "at random" via forum topics, a collaboration matchmaking tool holds significant potential.

The intersection of creativity and data science is an exciting frontier. We believe our methods have broader applications for analyzing and understanding complex social phenomena in a wide range of online communities. Let the fun begin!

References

[1] Newgrounds. http://www.newgrounds.com.

[2] Scratch: Programming for All. http://scratch.mit.edu.

[3] Slavin, R. Cooperative learning. *Review of Educational Research* 50, 2 (1980), 315–342.

[4] Burke, M., and Settles, B. Plugged in to the community: Social motivators in online goal-setting groups. In *Proceedings of the International Conference on Communities and Technologies (C&T)*, ACM (2011), 1–10.

[5] Lao, N., Mitchell, T., and Cohen, W. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL (2011)*, 529–539.

[6] Adamic, L., and Adar, E. Friends and neighbors on the Web. *Social Networks* 25, 3 (2003), 211–230.

[7] Liben-Nowell, D., and Kleinberg, J. The link prediction problem for social networks. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, ACM (2003), 556-559.

[8] McKenna, K., Green, A., and Gleason, M. Relationship formation on the Internet: What's the big attraction? *Journal of Social Issues* 58, 1 (2002), 9.

[9] Feld, S. The focused organization of social ties. *American Journal of Sociology* 86, 5 (1981), 1015–1035.

[10] Faulkner, R. R., and Anderson, A. B. Short-term projects and emergent careers: Evidence from Hollywood. *American Journal of Sociology* (1987). 879-909.

Biographies

Steven Dow is an assistant professor at the HCI Institute at Carnegie Mellon University where he researches human-computer interaction, design education, and social computing. He is co-recipient of two National Science Foundation grants, Stanford's Postdoctoral Research Award, and the Hasso Plattner Design Thinking Research Grant. He received an M.S. and Ph.D. in human-centered computing from the Georgia Institute of Technology, and a B.S. in industrial engineering from University of Iowa.

Burr Settles is a scientist and software engineer at Duolingo, a crowdsourcing ecosystem that combines foreign language education with translation. He also runs FAWM.ORG, an international songwriting challenge for musicians. He was previously a postdoc in machine learning at Carnegie Mellon University, and earned a Ph.D. in computer sciences from the University of Wisconsin-Madison. His book *Active Learning*—an introduction to learning algorithms that are adaptive, curious, and exploratory—was published by Morgan & Claypool in 2012. He has also co-organized several workshops on the subject at top international conferences.