Analysis of large-scale social and information networks

Jon Kleinberg

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

The growth of the Web has required us to think about the design of information systems in which large-scale computational and social feedback effects are simultaneously at work. At the same time, the data generated by Web-scale systems—recording the ways in which millions of participants create content, link information, form groups and communicate with one another—have made it possible to evaluate long-standing theories of social interaction, and to formulate new theories based on what we observe. These developments have created a new level of interaction between computing and the social sciences, enriching the perspectives of both of these disciplines. We discuss some of the observations, theories and conclusions that have grown from the study of Web-scale social interaction, focusing on issues including the mechanisms by which people join groups, the ways in which different groups are linked together in social networks and the interplay of positive and negative interactions in these networks.

1. Introduction

Understanding the World Wide Web requires us to consider not just the artefact itself, but also the set of larger-scale phenomena that it has catalyzed [1]. At the heart of these phenomena are a set of complex and overlapping networks, including the communication infrastructure on which the Web is constructed, the information networks that link its contents, and the social networks of interactions that it supports [2–4].

The study of the Web has, therefore, assumed a profoundly inter-disciplinary character, and in particular, has stimulated a growing dialogue between computing and the social sciences. This has turned out to be a very natural connection, for several reasons. First, the design
of computing systems in the era of the Web requires that we take into account not just technological constraints, but social ones as well. Patterns of social interaction form core components of modern information systems, and the performance of these systems depends on effective handling of social phenomena such as reputation, trust, status, incentives and social influence.

But there is more going on here than simply a broadening of the design space. As the Web has grown, we have been able to observe, measure and analyse social interaction at a level of scale and resolution that was previously unimaginable, observing social phenomena that had previously remained essentially unrecorded and invisible. These large-scale analyses are leading to new insights into long-standing questions in the social sciences, resulting in the refinement of existing theories and the formulation of new ones. Indeed, Web-scale social data are emerging as a natural counterpart to the type of data collected by more traditional social science methods, offering a complementary set of both advantages and challenges. The breadth and detail of Web data have made it possible to study rich phenomena that have been undetectable at smaller or coarser scales; but at the same time, this scale and lack of specificity has made it correspondingly more difficult to pose the kinds of complex questions that are possible when social science data are collected by targeted studies in more focused settings. The ongoing challenge is to bridge these gaps, bringing the two styles of research together in such a way that they both benefit from each other’s complementary strengths.

In this paper, we discuss recent studies that have used rich collections of data from the Web and online communication media to investigate fundamental questions concerning interaction in social networks. These questions include the tendency of people to form new social ties and affiliations; the contrasting roles played by different types of social ties; and the interplay between friendly and antagonistic relationships. In each case, we will focus on the flow of ideas between existing sociological theories and the insights and new theories that emerge from these large-scale analyses.

2. Forming ties and joining groups

Social networks, and social structures more generally, evolve as people develop new links to one another and establish affiliations with groups. There are a number of fundamental principles governing the tendency of these links and affiliations to form, and recent analyses of online data have quantified their effects across a range of different domains.

One of the basic principles at work is triadic closure [5,6], the increased tendency of two people to form a friendship when they have one or more friends in common. Kossinets & Watts [7] showed how this effect operates in the e-mail network among students at a large university, finding that common friendships make the formation of links more likely, and also analysing how this likelihood increases in the number of shared friends.

An equivalent way to formulate the principle of triadic closure is to say that people often form new links in a social network by connecting to someone two steps away from them—thus connecting to someone with whom they share a friend. With this view, we can refine the question of triadic closure by asking which two-step paths are most likely to support the formation of new social ties. Leskovec et al. [8] used data from a number of large social media sites to address this question; they employed maximum-likelihood methods to build models of link formation that took into account characteristics of the two-step paths along which new links tended to form. Among their other findings, they showed that two-step random walks are a simple but strong baseline predictor for the formation of new links: starting at a person A in the social network, they consider taking two random steps through the networks, and look at the probability of reaching person B as a proxy for the probability that A will form a link to B. Notice that if a node A has many neighbours in common with a node B, then this increases the chance that a two-step random walk from A will end at B, so the random-walk principle includes the notion that having many common neighbours increases the chance that a link forms. Their work showed that
further improvements in prediction were possible if additional features were taken into account, including the number of neighbours of nodes, as well as the level and recency of their activity in the system.

In addition to the formation of person-to-person links in a social network, affiliations also form between people and groups. One can view this as a link formation process in a larger network that includes both people and groups, with affiliation represented as the creation of a link between a person and a group [2,9,10]. Using data drawn both from social media and Web archives of professional collaboration, Backstrom et al. [11] studied some of the factors that increase a person’s tendency to join a group. They found that the probability a person joins a given group increases in the number of social ties the person has to people in the group. This effect generally shows a ‘diminishing returns’ property, in which successive links into the group have a decreasing effect. However, in a number of settings, there is a deviation from this diminishing returns behaviour at very small numbers of links—for example, in several domains, it is the case that having two links into a group can have more than twice the effect of having a single link.

Just as with triadic closure, one can ask more refined questions by looking carefully at the detailed network structure in the neighbourhood of a node. In particular, consider a person \( A \) who does not currently belong to a given group, but who has links to \( k \) members of the group—we can call these \( k \) people the neighbours of \( A \) in the group. An intriguing question is to look at how the number or density of connections among these \( k \) neighbours affects the probability that \( A \) joins the group [11]. In other words, does a highly interconnected set of neighbours exert more ‘gravitational pull’ on \( A \) compared with a sparsely interconnected set of neighbours? There are natural intuitive arguments in each direction, rooted in a well-articulated trade-off in sociology between the value of forming links to a highly connected set of people [12,13], and the value of forming links that span distinct, disconnected sets [6,14,15]. In the end, the direction of the effect likely depends on how a person construes the benefits of group membership in a particular domain: are they joining the group to become embedded in a tightly knit set of people (in which case, connections among friends in the group would be preferred), or are they joining to have access to disparate parts of their social neighbourhood, and potentially to bridge between these parts (in which case, a lack of connections among friends may be preferred)? In one particular setting, on LiveJournal, empirical studies showed that a person was more likely to join a group when their friends in the group were interconnected [11], but it is plausible that the direction of the effect is domain dependent. Ultimately, the main value for further investigation may lie in having the connectivity among friends in a group available as a significant parameter that can be varied in the analysis.

Finally, when we discuss the probability of a person joining a group, conditioned on \( k \) of their friends already having done so, there are complex questions of cause-and-effect at work. The analyses discussed thus far simply ask about the extent to which a set of social ties into a group serve as a predictor of future membership; the ways in which different underlying mechanisms actually facilitate membership are difficult to evaluate. In particular, there are likely several factors at work. First, if \( A \) has \( k \) friends in a group, this may well create a form of direct social influence leading \( A \) to join. But it is also the case that \( A \) is likely similar to her friends in interests and activities, and so the fact that several of these friends have already joined the group would be a useful predictor of \( A \)’s future behaviour, even if \( A \) were not aware of her friends’ presence in the group.

Sorting out the interplay of these factors is related to the complex issue of homophily in social networks—the fact that people are similar to those with whom they have social connections [16,17]. A basic question here is to separate out the dual forces by which similarity promotes friendship (because people seek out others like them) and by which friendship correspondingly promotes similarity (because people are influenced by their friends). Recent work has used Web data in which both relationships and behaviour can be observed over time to understand how changes in similarity and changes in social links interact [18–21]; this is an ongoing research challenge where these forms of time-resolved data at large scales have the potential to play a significant role.
3. Strong and weak ties

An important line of recent work using large-scale social data has tried to enrich the notion of a ‘link’ in the underlying social network. It is clear that the different social connections in a person’s life play different roles, and it is important to understand the distinctions among these roles both for better capturing the social phenomena at work, and also to better design new online applications.

Several pieces of empirical research using online data have addressed an influential theory in sociology owing to Granovetter [6], known as the strength of weak ties. Granovetter’s argument can be summarized in three parts. First, a person’s links in a social network can be divided into strong ties to close friends, which tend to be embedded in densely connected groups, and weak ties to more distant acquaintances. The links that span the boundaries of different groups tend to be weak ties. Second, these boundary-spanning weak ties are particularly crucial for conveying important information relevant to life changes and economic opportunities. Third, and as a consequence of the second point, people or communities with richer sets of boundary-spanning ties have more opportunities to be socially and economically successful. This last point has been developed in a rich line of work by researchers including Ron Burt, whose theory of structural holes articulates the advantages that come from bridging the gaps between disparate groups; Burt’s work has also offered empirical validation of these advantages in professional settings [14,15].

All three aspects of the strength of weak ties argument have been explored in recent work using large-scale data. Thus far, these studies have primarily employed digital records of communication data that exist in domains parallel to the Web itself—primarily e-mail and phone communication. However, it is clear that the methodology developed in these studies has close relevance to social interaction on the Web as well, and can be extended to this setting.

The first point in the argument has been explored by Onnela et al. in an analysis of cell-phone communication covering roughly 20 per cent of a national population [22]. They constructed a social network by defining links between people when they had engaged in phone communication, and they took the strength of the link between two people—the analogue of Granovetter’s division into strong and weak ties—to be proportional to the number of minutes the two people spent talking to each other by phone. In this way, the frequency of contact, which was available in their data, became a quantitative proxy for the harder-to-define notion of tie strength. They found that the strength of the link between two people was very strongly correlated with the overlap in these two people’s social neighbourhoods—in other words, with the number of common neighbours the two people had, as a proportion of their overall number of neighbours. This provides an important and large-scale validation of the premise that the links connecting people in disparate social circles tend to be weaker ties.

The second point in the argument, that a person’s access to boundary-spanning weak ties is important for providing them with new information, was initially developed by Granovetter in the context of his empirical study of job seekers in the Boston area. Adapting the question to the digital domain provides a good example of the challenges in moving to a large-scale online setting: with records of e-mail or phone communication, one has access to many more social interactions than one can obtain through direct interview methods, but it is correspondingly more difficult to see the subsequent consequences of these interactions. Is there a way to nonetheless quantify the ‘potential information gain’ from such social interactions?

Kossinets et al. [23] proposed an approach to this question by adapting the notion of vector clocks from the theory of distributed computing [24,25]. In effect, they measured the extent to which each person’s potential information about each other person in the social network was ‘out-of-date’ at any given point in time. When two people engage in a communication event, they then have the ability to update each other’s information about a possibly broad set of others to whom they have access—this is the potential information gain that results from the communication. On a large set of e-mail data, Kossinets et al. studied the effect of communication across links in the social network where the endpoints had no neighbours in common—in other words,
across local bridges that spanned groups. They found that on aggregate, communication across such boundary-spanning links produced greater potential information gain than communication events taking place on links where the endpoints had neighbours in common. While this is just a first, stylized measure of the potential ways in which information can move through social networks, it indicates how such notions of information flow can be formalized and compared across the different kinds of links that exist in a single network.

Finally, the third point in the argument—that access to boundary-spanning weak ties confers advantages—was investigated by Eagle et al. [26] using large-scale phone data from the UK. Their study considered a range of measures for the ‘diversity’ in a person’s social links, capturing the extent to which these links span different groups; they found that communities whose members have more diverse links tend to have correspondingly greater levels of economic development and economic opportunity, as measured by governmentally computed indices.

There is a remarkable level of complexity in the interaction among network structure, strengths of interpersonal contact, information flow and socioeconomic outcomes; and research on these issues is only beginning to understand how these relationships work. But there is considerable promise in the use of Web-scale data to provide fine-grained records of how these forces operate over time, at the level of large populations, combined with mathematical and computational models of the underlying effects and their consequences.

4. Positive and negative relationships

Social connections do not simply vary in whether they are strong or weak, they also vary in their polarity—whether they represent a friendly (positive) relationship or an antagonistic (negative) one. The interaction between positive and negative links in social networks has been the subject of a large body of theoretical research in sociology, beginning with the work of Heider, Cartwright and Harary in the 1940s and 1950s [27,28], but it is a topic where obtaining empirical data has been particularly challenging owing to the difficulty in reliably measuring the ‘sign’ of a link.

This too has been an area in which Web applications have provided rich data for studying how complex social relationships—in this case, positive and negative ones—interact at large scales. A number of Web sites have social features in which users can express both positive and negative sentiments toward others, and in clearly labelled ways that facilitate computational analysis. Leskovec et al. [29,30] analysed data from three such settings: the Epinions product review network, where users can express trust or distrust of each other; the technology blog Slashdot, where users can designate each other as ‘friends’ or ‘foes’; and a voting network among Wikipedia admins and other users, where individuals can vote positively or negatively on the promotion of others. Recent work of Szell et al. [31] looked at a further natural setting—the behaviour of players in a massively multi-player online game, where social interactions in the game include both cooperation and conflict.

In these domains, as in many Web-based settings, it is important to distinguish between analyses that treat the networks as undirected graphs (where a link represents a symmetric relationship between its endpoints) and those that treat them as directed graphs (where a link is viewed as going from one end to the other). In all four of the applications described above, the links are created by one user to point to another—either explicitly through a linking mechanism, as in Epinions and Slashdot, or implicitly through actions, as in Wikipedia voting or multi-player games.

There are two natural ways of interpreting a signed link from A to B. One interpretation is to view it as signifying a friend/enemy relationship; a positive link from A means ‘B is my friend’, whereas a negative link means ‘B is my enemy’. The theory of structural balance from social psychology [27,28] offers predictions for typical patterns of link signs in this case: they should reflect the notions that ‘the friend of my friend is my friend’, ‘the enemy of my friend is my enemy’, and more generally that A and B will be motivated to adjust their link signs so that they are friendly with each other precisely when they have the same friend/enemy status with respect to third parties C.
Another interpretation is to view a signed link as an expression of the relative status of $A$ and $B$, with $A$ indicating either deference or superiority through the sign of her link to $B$. Under this interpretation, a positive link from $A$ means ‘$B$ has higher status than I do’, whereas a negative link from $A$ means ‘$B$ has lower status than I do’. A companion theory of status, rooted in sociological theory and developed further through the study of directed linking on the Web, posits that in this case, signed links should approximately reflect relative levels in an underlying ordering of the nodes by status [29,32,33].

These competing interpretations are reflected in the patterns of signs among the links. For example, suppose $A$ links positively to $C$, and $B$ links negatively to $C$. Then, balance theory would argue that $B$ is the enemy of $A$’s friend, and hence would predict $A$ should link negatively to $B$. Status theory, on the other hand, would argue that $A$ looks up to $C$ while $B$ looks down on $C$; hence, $A$ has lower status than $B$, and should create a positive link to $B$. Leskovec et al. [29] show how contrasting predictions such as these can be used to assess, for a given signed network, whether balance or status is a better explanation for the aggregate pattern of link signs. For two datasets where sufficiently time-resolved data are available—Epinions and Wikipedia—they found that status is a better predictor of link signs than balance, although ingredients of both principles are clearly present. Moreover, techniques from machine learning can be used to provide predictions for the sign of a link from $A$ to $B$—based on the link signs they have expressed toward third parties—with an accuracy better than the predictions of either of the theories of balance or status [30].

These analyses of signed networks provide another dimension in which the analysis of online social interaction can be refined by a more detailed understanding of what the underlying links signify. There are many further directions in which these investigations can be broadened, and it is clear that the analysis of Web data will continue to provide us with new insights and new theories into the nature of social interaction in these large-scale settings.

The author is supported in part by a John D. and Catherine T. MacArthur Foundation Fellowship, a Google Research grant, a Yahoo! Research Alliance Grant and NSF grant nos IIS-0705774, IIS-0910664 and CCF-0910940.

References


