

Inferring Preference Correlations from Social Networks

Tad Hogg
HP Labs
Palo Alto, CA

April 27, 2009

Abstract

Identifying consumer preferences is a key challenge in customizing electronic commerce sites to individual users. The increasing availability of online social networks provides one approach to this problem: people linked in these networks often share preferences, allowing inference of interest in products based on knowledge of a consumer's network neighbors and their interests. This paper evaluates the benefits of inference from online social networks in two contexts: a random graph model and a web site allowing people to both express preferences and form distinct social and preference links. We determine conditions on network topology and preference correlations leading to extended clusters of people with similar interests. Knowledge of when such clusters occur improves the usefulness of social network-based inference for identifying products likely to interest consumers based on information from a few people in the network. Such estimates could help sellers design customized bundles of products and improve combinatorial auctions for complementary products.

Keywords: consumer preferences, social network homophily, network topology

1 Introduction

In social networks, people with similar characteristics tend to associate [49]. This property allows vendors to use available online social networks to plan targeted marketing via offline word of mouth interactions [18]. The similarity also enables social network administrators to improve collaborative filtering by weighting user preference correlations based on distance in their social network [42, 10, 76], and enhance online reputation mechanisms [31, 72] such as the one used on eBay [61].

Explicit or estimated knowledge of these networks [23] can also help vendors infer products or bundles of products likely to interest specific individuals based on observed behavior of a few people in the network. In particular, learning the preferences of one person allows inference of similar preferences for others nearby in the network [49], that is, people who are connected via a small number of links in the network. This use of social network structure allows:

- targeting a single product to members of the social network based on observations of others in the network, a form of collaborative filtering [42],
- identifying a bundle of products of likely interest to the group based on observations of interest in the bundle as a whole by some members of the network,
- suggesting possible bundles for further evaluation from observed preferences for individual products from several people who are close together in the network, and
- devising a product bundle for the collective consumption by the group, generalizing the concept of a bundle of products of interest to a single person.

Most discussion of product bundles focuses on purchases by individual consumers, primarily for complementary products. However, the last application in this list, identifying bundles for collective consumption, is especially beneficial for situations with allocative externality [39, 58], where the value of an item to an individual depends on what items others in the network have. An example is a vacation with friends having higher value for people in the network than each person taking the same vacation separately. The large datasets of shared preferences from online networks enable vendors to identify such groups with low cost, even when such situations are relatively rare.

Product bundles are usually created by the sellers, based on their expectations of how consumers value different combinations and the incremental cost of adding items to a bundle. With information goods, the marginal costs of additional items can be quite low (aside from possible licensing or royalty costs). Sellers will usually choose to offer only a few of the exponentially many possible bundles in the hope that these will be sufficient to allow price discrimination among consumers (e.g., between casual and dedicated users of a set of software tools).

If consumer preferences are sufficiently diverse or change rapidly, a fixed set of bundles will not efficiently capture the market potential. For example, the bundles may be priced

too high and thereby eliminate casual users from the market. On the other hand, low pricing may obtain high market share but may miss the potential revenue from dedicated users' higher values.

An alternative is to allow individual consumers to select bundles of interest. One approach to identifying high-value product bundles is a combinatorial auction [17]. These auctions accept bids on arbitrary bundles of the products for sale, thereby allowing bidders to express valuations for complementary and substitutable products (i.e., cases in which the value of the bundle is either greater or less than the sum of the individual item values). Even losing bids reveal consumer bundle preferences [37]. Unfortunately, combinatorial auctions involve considerable overhead to arrange and computational cost to run due to the difficulty of evaluating preferences over a potentially exponentially large number of combinations of bundles. Thus using combinatorial auctions to determine bundle preferences for large numbers of consumers and items can be prohibitive. An intermediate scenario is to allow consumers to select items with a simple bundle pricing scheme (e.g., a fixed price for any ten songs from a large set of songs). In this scheme of customized bundles [77], each consumer gains flexibility to select according to their individual interests, but without the complex overhead of adjusting prices for arbitrary bundles through a combinatorial auction.

Inferring joint preferences from available social networks is a low-cost method of estimating preferences, though less specific in eliciting preferences than combinatorial auctions. Thus network-based inference could be especially useful to identify situations where deploying a combinatorial auction could be worthwhile, and others where simpler bundling schemes likely suffice to capture much of the value. This means finding situations where there appear to be considerable correlation among preferences for items among niche subgroups of consumers. This inference relies on a correlation between a pair of users having a link and the similarity of their interests. Such correlations arise because people with similar interests tend to form links, or conversely because people who are linked influence the interests of each other. In either case, the observation of a link correlates with similar interests. This correlation is sufficient for the inference procedure described in this paper, which does not utilize more detailed models, such as causal mechanisms, leading to the observed correlations.

In the remainder of this paper, we describe methods for identifying products and bundles likely to interest particular individuals in the context of an available social network. We then illustrate the benefits of using preference correlations with a simple model of social networks, showing how the usefulness depends on the structure of the underlying graph and preference correlation in the graph. Finally, we examine the influence of link semantics using an online system allowing people to both express preferences and explicitly form links with differing nominal semantics. This example illustrates the significance for e-commerce applications of eliciting link semantics from users of social networks.

2 Identifying Preference Correlations

Online activities, such as search and web site visits, reveal users' interests in a readily recordable form. When aggregated over sites related to multiple products, the activity records can indicate possible interest in bundles of these goods or services.

Consider a data set of visits by many people to web sites relevant to various products or services. This data can include clicks on links, search keywords and time spent with various product web pages. One approach to identifying potential product bundles is through correlations in observed behaviors on sites relevant to those products. A variety of statistical techniques could identify significant correlations [63]. One example is modeling the effectiveness of search keyword advertising [25, 65]. A simple method looks for statistically significant deviations from expected values among all users, combined with any available prior knowledge of bundles purchased by others. Setting a detection threshold based on the observed distribution of user visits allows further detailed investigation to focus on cases with potential complementary interests to the user. This method generalizes to correlations among many web sites. Furthermore, instead of treating each site individually, we could group sites into various types (e.g., those involving automobiles or travel) and then look for correlations among two or more types that are significantly greater than would be expected for independent choices.

Using correlations to identify potential product bundles relies on probing individual behavior. This is a relatively slow process, requiring extensive records tracking single individuals as they use various web sites or search terms. Collaborative filtering addresses this limitation by inferring correlations based on other people with similar observed behaviors [27, 35], and can improve vendor performance [14, 22].

A key aspect of collaborative filtering is selecting which other users' behaviors to use to infer preferences of a particular user. Social networks directly created by users can provide strong correlations in addition to or instead of inferring correlation from commonly observed behaviors [42, 10, 76]. That is, the homophily of social networks reflects people with similar interests tending to link to each other [49]. Provided the links indeed reflect common interests relevant to a particular vendor or class of products, available online social networks can improve on inferring links based on behavior.

We can use these correlations in two ways. First, once a bundle of interest is identified for one person, we can infer that others in the network neighborhood are more likely than average to also have that preference. Second, people close together in a network who express interest in different products may also be more likely than average to have an interest in those products as a bundle. This inference, while not as strong as actually observing a bundle preference for each of the individuals, could substantially expand the usefulness of the network inferences. Specifically, instead of waiting for a few people to directly express their bundle preferences we can use the combined behavior of several individuals known to be relatively close in the network.

As vendors gain experience with inference accuracy, they can identify more specific measures related to the actual value of the bundles to consumers, such as actual purchase

histories. This procedure allows vendors, over time, to gain much of the same aggregate information as would be revealed through a series of combinatorial auctions, but without the cost of creating and running the auctions. A key question examined in the remainder of this paper is how network topology affects the usefulness of network-based inference, particularly the size of connected groups likely to have similar preferences.

3 Network Inference Performance

To illustrate the use of inference based on a social network, we consider a simple model as a theoretical benchmark for the empirical investigation discussed in Sec. 4. The social network among n people corresponds to a graph with n nodes. For simplicity, we consider undirected edges representing links between two people, without regard for the strength of the link. This simplification provides a useful model of social networks [52] despite ignoring properties such as asymmetric links and different types and strengths of interactions among individuals.

To model user preferences, consider a set of items for which each person has a binary preference, that is, either likes the item or does not. We divide the nodes of the graph into two groups: a fraction α representing people interested in the item, and the remaining nodes representing people who are not interested, as illustrated in Fig. 1. This formulation also applies to interest in specific subsets of items, corresponding to bundles of products. With this division of nodes, the graph contains three types of edges: those between two interested people, those between two uninterested people, and those between interested and uninterested persons.

Correlation of interests is reflected in the fractions of these three edge types in the graph. At one extreme, if linked people share the same interests, then the graph has no edges between interested and uninterested nodes. With no correlation in interests, an edge from an interested person randomly connects to any other person. Specifically, define ρ as the probability an edge from an interested person links to another interested person. Perfect correlation corresponds to $\rho = 1$, and no correlation corresponds to ρ equal to the fraction of available edges of each type from an interested person, that is, $\rho = \alpha$, on average.

3.1 Graph Properties

With this graph model, we can address a key issue for preference sharing: the size of connected components of interested people. Suppose the graph has degree distribution p_k , that is, a node has degree k with probability p_k , and aside from this constraint, edges are placed randomly. The subgraph consisting of just the interested people then has degree distribution

$$P_k = \sum_{j \geq k} p_j \text{Bi}(j, \rho; k) \tag{1}$$

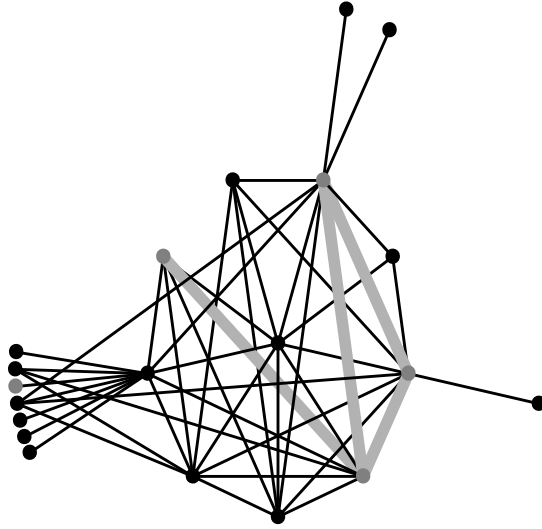


Figure 1: Example power-law graph showing the social network of a group of people. The gray nodes and links denote those interested in a particular item and links between interested people.

where $\text{Bi}(n, p; k) \equiv \binom{n}{k} p^k (1-p)^{n-k}$ is the binomial distribution, reflecting the likelihood k out of j links from a node will be to other interested people under the simplifying assumption that interests among neighbors are independent events.

The expected cluster size of this subgraph can be evaluated from the generating function of the degree sequence [53]

$$G(x) = \sum_k P_k x^k \quad (2)$$

Specifically, the expected number of neighbors of a node, z_1 , and the expected number of second neighbors, z_2 , are [53]

$$z_1 = G'(1) \quad (3)$$

$$z_2 = G''(1) \quad (4)$$

These values determine the size distribution of the components. Specifically, in the limit of large graphs (i.e., $n \rightarrow \infty$), when $z_1 > z_2$, the graph consists of relatively small components (i.e., of size $\ll n$) and when $z_1 < z_2$ most nodes are part of a single giant component.

3.2 Power-Law Graphs

Social networks have extended tails in their degree distributions, often well-described by power laws [8]. In practice, such power-law distributions apply only to a limited range of

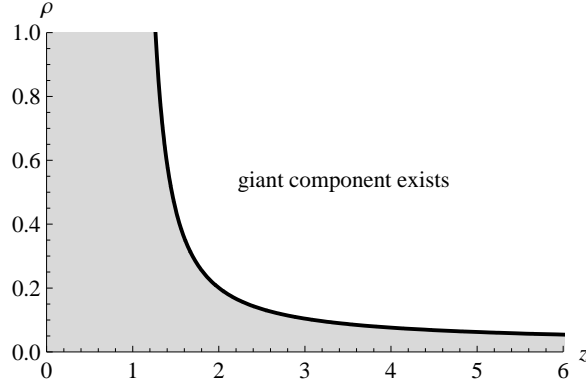


Figure 2: Behavior regimes for graph with power-law degree distribution with cutoff $\kappa = 20$. The giant component exists when the correlation among interested people ρ and average branching ratio of the full graph z are above and to the right of the curve. The lower right end of the curve has $\rho = 0.05$.

degrees, with more rapid decreases for larger degrees, which can arise when users become inactive, that is, no longer participate in the network [4]. To illustrate the behavioral regimes with the graph model described above, consider a graph with a truncated power-law degree distribution, namely

$$p_k \propto k^{-\tau} e^{-k/\kappa} \equiv k^{-\tau} \gamma^k, \quad (5)$$

for $k \geq 1$ with positive parameters τ and κ , and define $\gamma \equiv e^{-1/\kappa}$.

For the subgraph of interested people, the generating function $G(x)$ of Eq. (2) is

$$\frac{1}{\text{Li}_\tau(\gamma)} (\text{Li}_\tau(\gamma(1 - (1 - x)\rho)) - \text{Li}_\tau(\gamma(1 - \rho))) \quad (6)$$

where $\text{Li}_n(x)$ is the n th polylogarithm of x . With Eq. (3) and (4), this generating function gives the criterion for the appearance of the giant component, that is, $z_2/z_1 = 1$, in the large- n limit as

$$\rho \left(\frac{\text{Li}_{\tau-2}(\gamma)}{\text{Li}_{\tau-1}(\gamma)} - 1 \right) = 1 \quad (7)$$

Fig. 2 illustrates the resulting behavior for the degree distribution of Eq. (5) with $\kappa = 20$ and various values of τ giving a range of average degree z in the full graph. For large n , the curve in the figure indicates a threshold where the subgraph of interested people changes from having only relatively small clusters to having a giant component with a large fraction of all the interested people. Such thresholds are common in many graph models [47, 21, 74].

3.3 Cluster Size and Preference Correlations

The threshold for the appearance of the giant component in the subgraph of people interested in the item separates two regimes. In the first, *local*, regime, interested people are in small clusters within the social network. Thus when one interested individual is identified, the social network probably only allows inference of a few nearby individuals as significantly more likely than average to also be interested. The second, *global*, regime has an extended cluster of people with likely interest in the item. In this giant component, most interested people will not be direct neighbors of a single identified interested person.

This difference in connectivity among interested people affects strategies for spreading information about a product, for example, through coupons or relying on word-of-mouth recommendations. Specifically, these observations of graph structure regimes raise the practical question of how a vendor could find other interested people by exploiting preference correlations in the social network. A simple approach considers all people within a relatively small distance d of the identified individual. This is effective at identifying most of a small cluster. But in the regime with the giant component, this threshold method will exclude most of the cluster and hence miss many potential customers. In the global regime, vendors can use the social network to aid in identifying potential customers from among a large population, providing one approach to exploiting the long tail of preferences in economics [12, 5]. To improve the cluster identification in this case, one could actively probe for other members of the cluster and expand from them. Using the properties of power-law graphs improves the efficiency of such search procedures [3, 73], such as a focus on any hubs (i.e., high-degree nodes) near identified interested individuals. Alternatively, the global regime is well-suited for word-of-mouth spreading recommendations among interested people because members of the large cluster are reachable through links between interested people. Thus estimates of the global properties of the graph of interested people based on observed behavior of a few people and knowledge of their social network suggests appropriate strategy choices for e-commerce applications.

4 Example: Correlated Behavior in Essembly Networks

The graph model described above provides simple expressions for properties such as cluster size, but its use of random edge placement neglects some significant aspects of social networks. In particular, random edges do not correctly account for local clustering of social networks: two nodes linked to a third are significantly more likely to have a link between them than expected from random edge choices [74, 52]. Thus it is important to evaluate inference capabilities using more realistic networks. One approach involves graph models incorporating additional aspects of social interactions such as transitivity [41, 30, 62]. Another method, which we apply in this paper, is examining actual networks.

The rapid growth of user-contributory web sites provides a convenient source of large

networks [52]. These web sites rely on their users to provide and rate the web site’s contents. Examples include Digg, Flickr and YouTube, for news stories, photos and videos, respectively, as well as primarily social sites such as MySpace and Facebook. Many of these sites encourage users to form explicit links, indicating social relationships or shared interests. Alternatively, or in addition, the sites can provide *inferred* links among users based on common behaviors, such as a strong similarity in the contents viewed or rated. Such inferred links are a basis for collaborative filtering. Anonymized versions of the resulting networks and user activities are often available for academic research.

In the context of the preference inferences discussed in this paper, an ideal example would have a number of properties. First, the site would involve a large number of users, with many links and expressing preferences over many items. These large numbers allow evaluating the potential benefits of using the rapidly growing online networks in contrast to traditional survey studies, which can collect more detailed information on each person but are generally too expensive to apply to large groups. Second, the networks would be explicitly created by users to avoid the need to infer links based on somewhat arbitrary choices of sufficient common behavior to warrant a link. Third, the users could easily annotate links to differentiate among people they know socially and those they found on the web site. In the latter case, the decision to form a link accounts for whatever aspects of common behavior the users involved find significant. This distinction in link types would allow comparing the theory of Sec. 3 with semantically distinct networks. Fourth, the user activities on the site would indicate economically relevant behaviors, for example, willingness to purchase rated goods or services.

Many user-contributory sites have the first two properties: a large user population and explicitly created networks. Most of these sites have only a single “friend” link and so do not readily distinguish the semantics of the link. More significantly, the behaviors on these sites, such as contributing or rating news stories, do not specifically reveal commercial preferences. In contrast, the behavior and networks of commercial sites, such as Amazon or telephone calling patterns, are often proprietary or legally restricted and hence not readily available for large-scale study.

Because of the greater availability of data on noncommercial sites, and the similarity of the network structures among many such sites [52], in this paper we consider an example with several user-created networks, namely Essembly (www.essembly.com), a social network facilitating the formation of ideological political groups. Members post resolves reflecting controversial opinions which others can then vote and comment on. Example resolves are “overall, free trade is good for American workers” and “all speech – even the most offensive – should be protected under the First Amendment”.

Essembly offers both a social network (friends) and ideological networks (allies and nemeses). Nominally, links in the friends network connect people who know each other, while those in the allies and nemeses networks link people who tend to agree or disagree, respectively, about the political issues important to the users. Links in these three networks are explicitly created by the users, and approved by both users involved in the link. That is, the users themselves mutually choose whom to link to (and whether to

connect as friend, ally or nemesis). The aggregate user behavior and network properties of Essembly [34] are similar to those of other user-contributory web sites, such as Digg and Flickr [43, 75]. But unlike these other sites, with a single “friend” link, Essembly offers distinct links for social and preference relationships among users.

With explicit link creation, Essembly avoids the question of whether inferences based on behavior correctly represent shared preferences. This contrasts with links inferred from observed similar user behavior as used with collaborative filtering. For example, users may consider a few issues as extremely important in identifying people with shared interests, while inferred links could miss this if those users happen to have many shared behaviors on issues they regard as minor. For example, two users with common votes on a few resolves on environmental issues but many different votes on food or pet preferences may decide to link themselves as allies, whereas inference based on all the resolves these users have both voted on will not identify this commonality.

The Essembly dataset examined here consists of anonymized voting records and network connections from August 2005 through December 2006. The data include 15,419 users, 24,963 resolves and about 1.4 million votes. The Essembly site used 10 of these resolves during user registration to create an initial ideological profile of the user. These 10 resolves have far more votes than the others. The examples discussed in the remainder of this section use only the remaining, that is, user-created, resolves. These resolves account for about 1.3 million votes.

The Essembly networks appear to follow their nominal semantics [34]. In particular, allies tend to vote the same on many resolves while nemeses vote oppositely. Friends are intermediate between allies and nemeses. Moreover, all three networks show a truncated power-law degree sequence as in the model described above [34]. The actual networks are somewhat more clustered than the random edges assumed in the model, as expected with social networks [52]. Clustering is an example of the effect of weak ties in networks [28].

Essembly users do not express valuations for resolves or groups of resolves, so the Essembly data cannot directly address inference of complementary *values* for bundles. In this respect, Essembly is similar to the other social web sites mentioned above. Essembly nevertheless allows comparing *preference* correlations in three distinct *user-created* networks in a simplified context of binary preferences. This example thus provides some indication of the usefulness of networks for inference of joint preferences among users of real online networks. Moreover, for some practical scenarios in which sellers create bundles, identifying items consumers have some interest in can be useful in itself even without knowledge of *how much* consumers value those items [77].

4.1 Local Preference Correlation and the Giant Component

A simple proxy for interest in a resolve is a user’s choice to vote on that resolve. The correlation in behavior among users in the model described in Sec. 3, that is, the value of ρ , corresponds to correlation in votes on resolves. Similarly, the fraction of users interested, that is, α , is the fraction of users who voted on that resolve. For Essembly users, ρ tends to

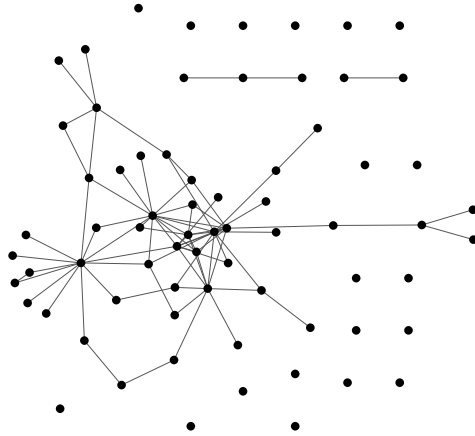


Figure 3: Subgraph of the Essembly friends network consisting of the 70 voters on a single resolve. This case has $\alpha = 0.014$ and $\rho = 0.067$, that is, 1.4% of the members of the friends network voted on this resolve, and 6.7% of a voter’s neighbors in the network also voted on the resolve, on average.

be larger than α , most notably in the friends network, reflecting homophily in the networks¹.

Fig. 3 shows the components of the subgraph of voters in the friends network for a particular resolve. In this case, about 60% of the voters are in the largest component. In addition to the aggregate measures of ρ and α , this figure shows considerable variation in the local clustering of voters in the subgraph, reflecting the large variance in numbers of neighbors typical of social network graphs [52].

Fig. 4 uses a sample of resolves to show how correlation in behavior relates to the size of the giant component, in the subgraph of the network containing the voters on each resolve. For comparison, the dashed line is the threshold for the appearance of a giant component in the random power-law graph model using the truncated power-law fit to the degree sequence of each network [34]. Although the threshold strictly only applies in the limit of large n , for the preference networks in Essembly, the threshold gives a reasonable indication of whether the giant component is likely to contain most of the voters on a resolve. Thus the random graph model allows estimating the likely extent of the largest cluster based on observed preference correlations among neighbors of a few people who express interest in a resolve. This model is less informative for distinguishing cluster size for the friends network as the largest component size shows a gradual increase with preference correlation ρ .

A second observation from Fig. 4 is the *semantics* of the links in a network affect the

¹There are a few cases of anticorrelation (ρ less than α) in the nemeses network.

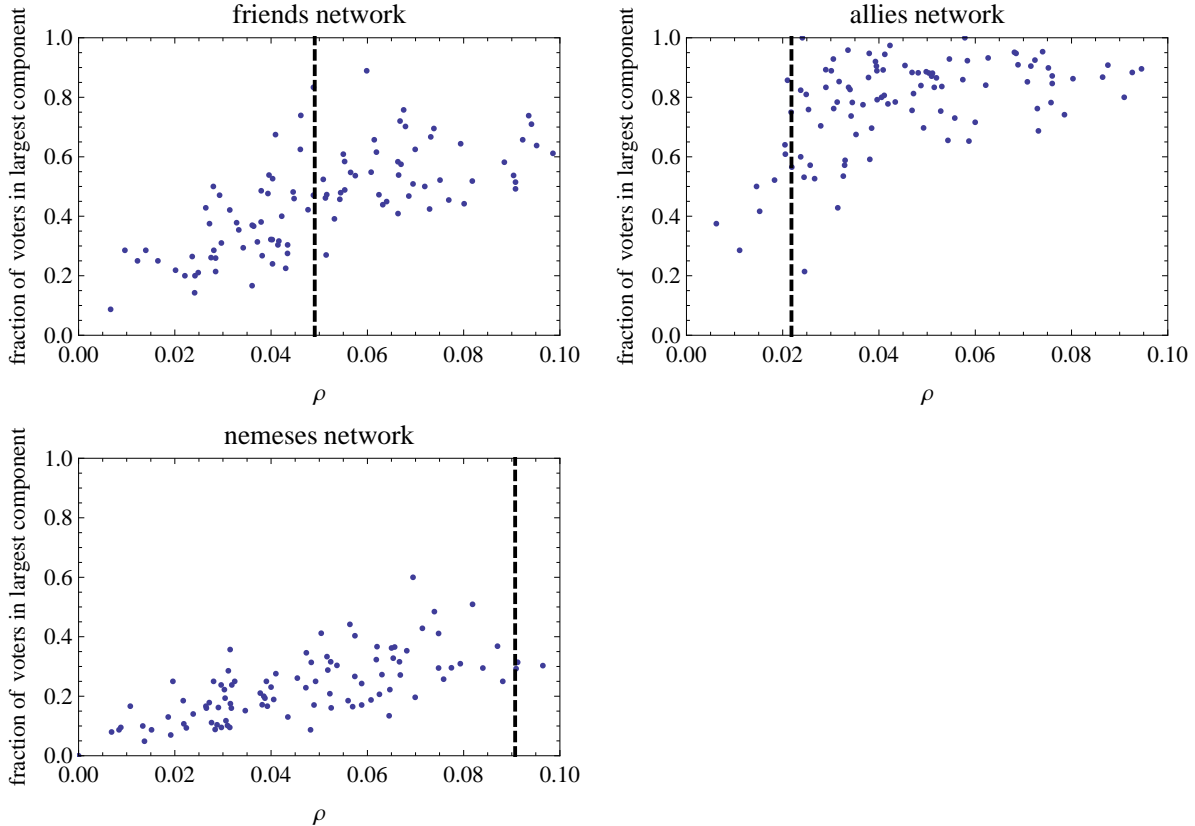


Figure 4: Fraction of voters in largest component vs. local correlation ρ in the subgraphs of voters for a sample of 100 resolves. Each point represents a distinct resolve, and the same set of resolves appears in the plot for each network. The dashed line in each plot indicates the predicted threshold, from Eq. (7), above which the giant component appears in the random graph model with many nodes.

cluster size, indicating the importance for e-commerce applications of eliciting the type of relationship a link entails, which is not made explicit on most social network sites [34]. Thus depending on the application, a vendor could select a type of link more or less likely to give large clusters of consumers with similar preferences for particular products or bundles. Large clusters could facilitate word-of-mouth marketing while small clusters could be useful for surveys to sample a variety of opinions less likely to influence each other. The latter is particularly relevant in using social networks since such sites often promote users learning about others' activities and preferences via their links in the network.

4.2 Inference of Joint Preferences

A key use for networks is inference of common interests among users. The Essembly dataset allows the quantification of this inference by counting resolves voted on by pairs of

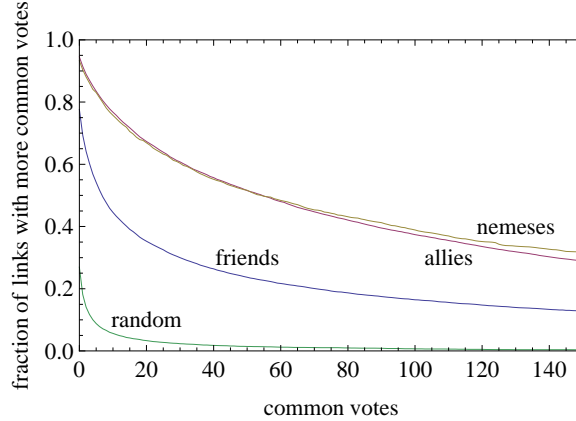


Figure 5: Cumulative distribution of number of common votes among pairs in the networks, and among random pairs of users. For each number of common votes, the curves show the fraction of pairs with more than that many resolves both users in the pair voted on.

users linked in the networks. For single item inference, in any of the three networks, if two people have a link and one is known to have voted on a resolve, the other has about a 20% chance of having voted on it, compared with 4% for random pairs of users. These values represent the average over the pairs of users of the ratio of the number of resolves they both voted on to the number voted on by the first member of the pair. As an example of bundle inference using information from different people, consider two people, each of whom is known to have voted on a different resolve. In this situation, the chance that both people voted on both resolves is about 1.3% if they share a link, compared to 0.06% for random pairs of users. The probabilities decrease further for larger bundles. For example, if three people are each known to have voted on a different resolve, the probability they all voted on all three of those resolves is about 10^{-4} when those three people are linked in one of the Essembly networks (i.e., form a triangle subgraph of the network). On the other hand, for random triples of users the probability is only about 10^{-8} .

For these examples, randomization tests [15] indicate the inference probabilities for users with a network link are unlikely to be the same as those from random pairs of users, with p -value less than 10^{-3} . For these inference examples, the results are similar among the three networks. While the probabilities are small, especially when considering bundles with more than two items, the likelihood of successful inference using network links is significantly improved over random users.

For further insight on inference performance, Fig. 5 illustrates a distinction between the ideological networks in Essembly and the social network nominally linking people who know each other as friends. The figure shows friends generally have many more resolves in common, that is, resolves both users voted on, than random pairs of users. This underlies the ability to infer joint preferences and form large clusters of interested users based on a social network. The figure also shows the preference networks (both allies and nemeses) are

similar and have significantly more common resolves than the friends network. Essembly presents ideological networks as opportunities for users to link to others based on preferences exhibited in Essembly rather than via personal familiarity. This situation is analogous to online product recommendation systems, where people rely on reviews from others they do not know socially.

The larger number of common resolves in the ideological networks suggests e-commerce sites could facilitate inference of likely product interests by encouraging users to explicitly specify links to others who they deem share their interests, as distinct from only having a single “friends” network nominally intended to link people who already know each other. In practice, users could learn of such common interests from publicly disclosed behavior of other users on a vendor’s web site or a site devoted to reviews of particular types of products. Such information is valuable in its own right in helping users make purchase decisions [20]. In conjunction with feedback in online reputation systems [60] and aids to propagate trust through user-specified links [26, 29, 71], this information on common interests could be extended to a wider range of users. That is, users could reveal their identification of others with similar (or dissimilar) interests by forming links, in analogy with the ideological networks of Essembly, if the e-commerce site allows users to create distinct link types. The addition of various link types may also help reduce the dilution of interest correlations among linked users if people feel pressured to accept link requests from others they do not know well. Instead they could select link types that distinguish people they know well from others with shared interests.

5 Discussion

This paper examined the use of preference correlation in an analytically tractable model of social networks and one online community, Essembly. It would be useful to compare the graph model examined here with publicly available data from web sites with both product purchases and networks with users explicitly indicating link semantics (e.g., a social link to friends vs. a similarity link to people with similar interests). Such data would allow estimating valuations for inferred bundles of products. Moreover, while social networks from a variety of contexts have similar structural features [52], there remains the question of whether common interests seen on sites such as Essembly, Digg, Flickr and YouTube have similar inference power with networks for product purchases.

Inference with social networks is particularly useful for products with low marginal production cost, such as information goods. In that case, customizing the product for a group with similar preferences aggregates many transactions without much additional cost to discover the group. This customization can be profitable even when a series of separate transactions would not be (e.g., due to the repeated transaction costs for low-margin products). As another application, estimates of preference correlations could allow gerrymandering subgroups selected for surveys or spreading recommendations, for example, distributing or concentrating preferences in selected subgroups so the majority of group reports would be opposite that of the majority opinion of people in the network. Such

procedures can affect group choices through designing the decision procedure [56] or aid preference elicitation [70].

The network used for inference need not just arise from online behavior or explicit user-created links. Instead, links could also represent observed correlated physical behaviors (e.g., from cell phone locations or radio frequency identifier tags (RFIDs)) or existing demographic data [46, 55, 76]. Furthermore, inference could depend on the strength and type of commonality represented by a link [23], for example, preferences for items (based on similarity of purchase history), or commonality of trust, incentives or reputations about the items, the vendor or other people. For example, online reputation systems often provide a variety of information (e.g., a numerical rating, text comments and relationships to other users via networks) with differing relevance for users [24, 72]. This multiplicity of link types contrasts with most online social networks which collapse these relationships into a single unweighted link type.

Estimating preferences by inference through social networks takes more time to accumulate statistical significance than direct measures from a preexisting active market. Thus using social networks may not identify transient interest in products (e.g., due to a breaking news story or sports event) unless there is a corresponding rapid increase in online user activity (e.g., via blogs) [45]. This inference delay raises challenges in using information to make decisions. Delays can lead to complex dynamics [32] such as transients or oscillations, potentially leading firms to create customized products or bundles after they are no longer of interest. Nevertheless, using online behavior with social networks can respond far more rapidly to unexpected changing preferences than traditional methods such as market surveys. More importantly, network-based inference can identify potential combinations of products with high interest for which no survey or specific market yet exists.

Inferences based on shared preferences could facilitate the use of product bundles in e-commerce: prior inference of likely bundles for nearby people in the network suggests customized bundles for sellers to offer [77]. The inference could also reduce the search for winner(s) in combinatorial auctions. Such inference, while not always accurate, complements other proposals to reduce search time in combinatorial auctions by restricting the allowed bids [64, 69, 54, 16]. Inferences could also facilitate the use of coupons or other price discrimination [2, 6] to network neighbors of those who have previously purchased similar goods.

A significant challenge for inference with networks is the available information is only an approximation of people's relationships and preferences. For use of social networks in e-commerce, the network information could be incomplete and out of date, that is, noisy. Thus in practice, evaluating the usefulness of network-based inference for e-commerce requires understanding the consequence of errors in the data. Fortunately, mechanisms relying on aggregated information from social networks are somewhat robust: performance degrades gradually rather than abruptly with noise [33]. In such cases, estimates of consumer interests based on approximate network information is beneficial compared to not using the information at all. Evaluating the amount of noise in online networks and its

effects on mechanisms relying on those networks is an important direction for future work. One approach to noise is using robust graph properties, that is, those whose values don't change much if graph is perturbed. For example, one distance in social networks is the shortest path in the graph between two nodes. But a weighted average of the different paths may correspond more realistically to social network influence, that is, if I have many friends in common with a friend-of-a-friend, preferences may be more correlated than if we only share one linking friend. Such effects can be included in more complex graphical models of social networks [62] than those described in Sec. 3. An active approach to reducing noise is using a mechanism to induce participants to truthfully reveal their links, for example by rewarding them for estimating the behavior of others in their network, as previously proposed for unstructured groups [13, 57].

Another challenge in using social networks, as well as inferred similarities among users from their behavior, is privacy concerns over how the information may be misused. These concerns may arise from purchases that violate norms of a user's social group. In addition, hiding such information can be important for purely economic reasons to avoid price discrimination [1]. Approaches to this challenging problem [51] include cryptographic techniques to provide matching while maintaining privacy [36, 50, 59] and public policy safeguards [9, 40].

A further challenge arises from the feedback of using networks in e-commerce on the incentives users have for forming links. Vendors' use of networks may alter the terms of e-commerce transactions whose combined value to users becomes larger than the value they gain from accurate links. If so, users have incentives to misrepresent their links to gain more favorable terms. This possibility requires designing the overall use of the networks to encourage users to provide accurate links. Models of network formation including the costs and benefits users obtain from links [38] are one approach to developing such designs.

Addressing many of these challenges would benefit from a predictive understanding of how and when social networks link people who influence each others' preferences, rather than just reflecting prior similarities. Vendors could use such knowledge to identify situations where new preference correlations might be created through a social network, such as with viral marketing [44]. While available online networks can include thousands or millions of users, and thus give strong statistical correlations, detailed information on why users form links is usually lacking. Thus it is difficult to distinguish links arising from prior similarity from influence of linked individuals creating similar preferences [48]. Surveys can provide such information, but are expensive. Thus, as with the example in this paper, most such studies are limited to available observational data. In such cases, even strong statistical correlations among behaviors cannot make definite causal predictions, such as how people will change their purchase decisions if a vendor provides discount coupons to network neighbors. This restriction arises from the possibility of unobserved effects missing from the data or selection effects of which users chose to participate [48, 19].

One approach to identifying likely causal influences is fitting plausible general social mechanisms for network formation to observations of link creation over time [68]. Statistical estimation methods can then suggest which observed correlations indicate likely

causal mechanisms, for example as influencing consumer choices [11]. More precisely, such studies *rule out* cases where correlation could be causal, for example, by testing for inconsistent temporal relations between one change, such as link formation, and another, such as a new preference revealed by user behavior. These approaches focus on giving plausible mechanistic causal explanations for observed statistical correlations. When such explanations are sufficiently well-specified to allow a computational implementation [19], the resulting algorithms allow vendors to use these causal models to estimate how customer preferences may change through network influence. In practice, such computational models could improve vendor outcomes even if they only approximately characterize causal influences.

These approaches to causal models, based on observational data, can be misleading when other, unobserved, influences account for the correlations. Intervention experiments can provide more definitive understanding of causation. That is, deliberately changing one variable in a controlled setting and seeing which other variables also change. Such experiments face a trade-off among cost, control of extraneous variables and scope in terms of number of users and time they are observed. At one extreme, vendors who act based on correlations, for example, to introduce new product bundles or offer purchase incentives to highly connected people in a network, are performing real-world experiments but with little control over the user incentives and other influences. Laboratory economic experiments [67] are the other extreme, allowing control over users' incentives and their communication. However, such experiments are limited to relatively small groups of people and behavior over short times, typically a few hours. The recent development of web-based experiments [66] and online games [7] involving economic transactions provide far larger groups than possible in laboratory settings, though with less control over user incentives and interactions. The differing trade-offs involved in these approaches are complementary, allowing testing of hypothesized causal mechanisms in a variety of scenarios, ranging from well-controlled but unrealistic to less controlled real-world practice. Such experimental studies could lead to better identification of underlying causal mechanisms. Knowledge of such mechanisms could improve use of social networks in e-commerce, particularly involving niche correlations among user preferences for a variety of information goods and services.

Acknowledgments

I thank Chris Chan and Jimmy Kittiyachavalit of Essembly for providing the Essembly user data, and Michael Brzozowski, Bernardo Huberman and Dennis Wilkinson for helpful discussions.

References

- [1] Alessandro Acquisti. Identity management, privacy, and price discrimination. *IEEE Security and Privacy*, 6(2):46–50, 2008.

- [2] Alessandro Acquisti and Hal R. Varian. Conditioning prices on purchase history. *Marketing Science*, 24:367–381, 2005.
- [3] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Physical Review E*, 64:46135, 2001.
- [4] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proc. of the Natl. Acad. Sci.*, 97:11149–11152, 2000.
- [5] Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 2006.
- [6] Joseph P. Bailey. Internet price discrimination: Self-regulation, public policy, and global electronic commerce. Technical report, School of Business, University of Maryland, College Park, MD, 1998.
- [7] William Sims Bainbridge. The scientific research potential of virtual worlds. *Science*, 317:472–476, 2007.
- [8] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [9] Ronald Bayer and Amy L. Fairchild. Surveillance and privacy. *Science*, 290:1898–1899, 2000.
- [10] David Ben-Shimon et al. Recommender system from personal social networks. In K. M. Wegrzyn-Wolska and P. S. Szczepaniak, editors, *Advances in Intelligent Web Mastering*, pages 47–55. Springer, Berlin, 2007.
- [11] Daniel Birke. Who you are or whom you know? Consumption interdependences in social networks. In *Proc. of the Royal Economic Society Annual Conference*, March 2008.
- [12] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D. Smith. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49:1580–1596, 2003.
- [13] Kay-Yut Chen, Leslie R. Fine, and Bernardo A. Huberman. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50:983–994, 2004.
- [14] Pei-Yu Chen and Shin-yi Wu. Does collaborative filtering technology impact sales? empirical evidence from Amazon.com. Working Paper 1002698, Social Science Research Network, July 2007.
- [15] Paul R. Cohen. *Empirical Methods for Artificial Intelligence*. MIT Press, Cambridge, MA, 1995.
- [16] Vincent Conitzer, Jonathan Derryberry, and Tuomas Sandholm. Combinatorial auctions with structured item graphs. In *Proc. of the 19th Natl. Conf. on Artificial Intelligence (AAAI2004)*, pages 212–218, 2004.
- [17] Peter Cramton, Yoav Shoham, and Richard Steinberg, editors. *Combinatorial Auctions*. MIT Press, Cambridge, MA, 2006.

- [18] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proc. of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 57–66. ACM Press, 2001.
- [19] Patrick Doreian. Causality in social network analysis. *Sociological Methods & Research*, 30:81–114, 2001.
- [20] Chris Forman, Anindya Ghose, and Batia Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3), 2008.
- [21] Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proc. of the American Mathematical Society*, 124(10):2993–3002, 1996.
- [22] Robert Garfinkel et al. Empirical analysis of the business value of recommender systems. Working Paper 958770, Social Science Research Network, November 2006.
- [23] Laura Garton, Caroline Haythornthwaite, and Barry Wellman. Studying online social networks. *J. of Computer Mediated Communication*, 3(1), June 1997.
- [24] Anindya Ghose, Panagiotis G. Ipeirotis, and Arun Sundararajan. Reputation premiums in electronic peer-to-peer markets: analyzing textual feedback and network structure. In *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, pages 150–154, NY, 2005. ACM.
- [25] Anindya Ghose and Sha Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. NET Institute Working Paper 07-35, Social Science Research Network, Sept. 2007.
- [26] Jennifer Globeck. Weaving a web of trust. *Science*, 321:1640–1641, 2008.
- [27] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.
- [28] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1:201–233, 1983.
- [29] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proc. of the 13th Intl. World Wide Web Conf. (WWW2004)*, pages 403–412, New York, 2004. ACM.
- [30] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *J. of the American Statistical Association*, 97:1090–1098, 2002.
- [31] Tad Hogg and Lada Adamic. Enhancing reputation mechanisms via online social networks. In *Proc. of the 5th ACM Conference on Electronic Commerce (EC'04)*, pages 236–237. ACM Press, 2004.
- [32] Tad Hogg and Bernardo A. Huberman. Dynamics of large autonomous computational systems. In Kagan Tumer and David Wolpert, editors, *Collectives and the Design of Complex Systems*, pages 295–315. Springer, New York, 2004.

- [33] Tad Hogg and Bernardo A. Huberman. Solving the organizational free riding problem with social networks. In K. Lerman et al., editors, *Proc. of the AAAI Symposium on Social Information Processing*, pages 24–29, 2008.
- [34] Tad Hogg, Dennis M Wilkinson, Gabor Szabo, and Michael Brzozowski. Multiple relationship types in online communities and social networks. In K. Lerman et al., editors, *Proc. of the AAAI Symposium on Social Information Processing*, pages 30–35, 2008.
- [35] Zan Huang, Daniel Zeng, and Hsinchun Chen. A comparative study of recommendation algorithms in e-commerce applications. *IEEE Intelligent Systems*, 22:68–78, 2007.
- [36] Bernardo A. Huberman, Matt Franklin, and Tad Hogg. Enhancing privacy and trust in electronic communities. In *Proc. of the ACM Conference on Electronic Commerce (EC99)*, pages 78–86, NY, 1999. ACM Press.
- [37] Bernardo A. Huberman, Tad Hogg, and Arun Swami. Using unsuccessful auction bids to identify latent demand. In *Proc. of the IEEE Conference on Systems, Man and Cybernetics*, pages 2911–2916, 2001.
- [38] Matthew O. Jackson. A survey of models of network formation: Stability and efficiency. In G. Demange and M. Wooders, editors, *Group Formation in Economics: Networks, Clubs and Coalitions*, chapter 1. Cambridge Univ. Press, Cambridge, 2004.
- [39] Philippe Jehiel and Benny Moldovanu. Allocative and informational externalities in auctions and related mechanisms. Technical Report SFB/TR 15 142, Free University of Berlin, 2005. available at ideas.repec.org/p/trf/wpaper/142.html.
- [40] Alden S. Klovdahl. Social network research and human subjects protection: Towards more effective infectious disease control. *Social Networks*, 27:119–137, 2005.
- [41] Johan H. Koskinen and Tom A. B. Snijders. Bayesian inference for dynamic social network data. *J. of Statistical Planning and Inference*, 137:3930–3938, 2007.
- [42] Chuck Lam. SNACK: incorporating social network information in automated collaborative filtering. In *Proc. of the 5th ACM Conference on Electronic Commerce (EC’04)*, pages 254–255. ACM Press, 2004.
- [43] Kristina Lerman. User participation in social media: Digg study. In *IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology*, pages 255–258, 2007.
- [44] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), 2007.
- [45] Jure Leskovec et al. Cost-effective outbreak detection in networks. In P. Berkhin et al., editors, *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining (KDD07)*, pages 420–429. ACM, 2007.
- [46] David Liben-Nowell et al. Geographic routing in social networks. *Proc. of the Natl. Acad. of Sciences USA*, 102:11623–11628, 2005.

- [47] Tomasz Luczak. Phase transition phenomena in random discrete structures. *Discrete Mathematics*, 136:225–242, 1994.
- [48] Charles F. Manski. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60:531–542, 1993.
- [49] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
- [50] Moni Naor, Benny Pinkas, and Reuben Sumner. Privacy preserving auctions and mechanism design. In *Proc. of the ACM Conference on Electronic Commerce (EC99)*, pages 129–139, NY, 1999. ACM Press.
- [51] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proc. of the 30th IEEE Symposium on Security and Privacy*, 2009.
- [52] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
- [53] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.
- [54] David C. Parkes and Lyle H. Ungar. Iterative combinatorial auctions: Theory and practice. In *Proc. of the 17th Natl. Conf. on Artificial Intelligence (AAAI2000)*, pages 74–81. AAAI, 2000.
- [55] Alex (Sandy) Pentland. Automatic mapping and modeling of human networks. *Physica A*, 378:59–67, 2007.
- [56] Charles R. Plott and Michael E. Levine. A model of agenda influence on committee decisions. *American Economic Review*, 68:146–160, 1978.
- [57] Drazen Prelec. A Bayesian truth serum for subjective data. *Science*, 306:462–466, 2004.
- [58] Martin Ranger. The generalized ascending proxy auction in the presence of externalities. Working Paper 834785, Social Science Research Network, July 2005.
- [59] Michael K. Reiter and Aviel D. Rubin. Anonymous web transactions with crowds. *Communications of the ACM*, 42(2):32–38, February 1999.
- [60] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Communications of the ACM*, 43:45–48, 2000.
- [61] Paul Resnick, Richard Zechhauser, John Swanson, and Kate Lockwood. The value of reputation on eBay: A controlled experiment. *Experimental Economics*, 9:79–101, 2006.
- [62] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2007.

- [63] Peter E. Rossi and Greg M. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22:304–328, 2003.
- [64] Michael H. Rothkopf, Aleksandar Pekec, and Ronald M. Harstad. Computationally manageable combinatorial auctions. *Management Science*, 44:1131–1147, 1998.
- [65] Oliver J. Rutz and Randolph E. Bucklin. A model of individual keyword performance in paid search advertising. Working Paper 1024765, Social Science Research Network, June 2007.
- [66] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854–856, 2006.
- [67] Vernon L. Smith. *Bargaining and Market Behavior: Essays in Experimental Economics*. Cambridge Univ. Press, 2000.
- [68] Tom A. B. Snijders, Christian E. G. Steglich, and Michael Schweinberger. Modeling the co-evolution of networks and behavior. In K. van Montfort et al., editors, *Longitudinal Models in the Behavioral and Related Sciences*, pages 41–71. Lawrence Erlbaum, 2007.
- [69] Moshe Tennenholtz. Some tractable combinatorial auctions. In *Proc. of the 17th Natl. Conf. on Artificial Intelligence (AAAI2000)*, pages 98–103. AAAI, 2000.
- [70] Toby Walsh. Representing and reasoning with preferences. *AI Magazine*, 28(4):59–69, 2007.
- [71] Frank E. Walter, Stefano Battiston, and Frank Schweitzer. A model of a trust-based recommendation system of a social network. *Autonomous Agents and Multi-Agent Systems*, 16:57–74, 2008.
- [72] Jyun-Cheng Wang and Chui-Chen Chiu. Recommending trusted online auction sellers using social network analysis. *Expert Systems with Applications*, 34:1666–1679, 2008.
- [73] Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.
- [74] Duncan J. Watts and Steven H. Strogatz. The importance of being connected: Structure, dynamics and games in a small world. Technical report, Dept. of Theoretical and Applied Mechanics, Cornell Univ., 1997.
- [75] Dennis M. Wilkinson. Strong regularities in online peer production. In *Proc. of the 2008 ACM Conference on E-Commerce*, pages 302–309, 2008.
- [76] Wolfgang Woerndl and Georg Groh. Utilizing physical and social context to improve recommender systems. In *Proc. of the IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology Workshops*, pages 123–128, 2007.
- [77] Shin-yi Wu, Lorin M. Hitt, Pei-yu Chen, and G. Anandalingam. Customized bundle pricing for information goods: A nonlinear mixed-integer programming approach. *Management Science*, 54:608–622, 2008.