CS 224w - Reaction Paper

Alexander Loewi, Jan Overgoor and Evan Rosen

October 6, 2010

We are interested in applying decentralized search algorithms to real world high school data in which individuals exist simultaneously in several social networks. We have access to a data set containing the interactions of students in high school class rooms along with a detailed profile for each student. We wish to investigate the searchability of various networks which can be constructed from these data. In particular, we are interested in the way that several superimposed networks, capturing different aspects of student interaction, might effect search in such an environment. In this paper we therefore consider both sociological research with application in network theory as well as literature on decentralized search.

1 Social Science

McFarland and Thomas's 2006 paper "Bowling Youth" [McFarland and Thomas, 2006] collects data of various measures of political engagement—such as voting, involvement in a campaign, holding a position in a political organization—and correlates them to a host of data on high school experiences. The focus on young adults is cited as relevant due to the trend of political engagement seeming to stay mostly constant in adulthood, thus making high school potentially a particularly important time for influencing life long engagement. While it is not the only thing measured (standard socioeconomic factors are also taken into account), student involvement in clubs and organizations within the school are the main focus of the analysis. The reason for this is that the work builds off of the book *Bowling Alone* [Putnam, 2000] which correlates what is deemed *social capital* to a large number of varied trends of which civic engagement is only one. However as "social capital" is measured in *Bowling Alone* largely as a function of the organizations to which a population belongs, the logical path from group membership to political involvement is a clear one. The paper finds a large number of groups, notably and mostly predictably service clubs, student council, but also drama clubs, all have modest but significant effects that can be seen to last between six and twelve years.

While the paper does find a large number of interesting correlations, it quite bizarrely does not focus on or even point what it identifies as the most strongest factors, but rather emphasizes the general importance of groups in general. (As

an interesting note, the only thing more negatively correlated with engagement than computer club was field hockey.) The most positively correlated factors were the schools' being private (and the number of clubs a school offers, which we think are very closely related) and membership in the National Honors Society, which strongly outperform every other factor. It was previously hypothesized [Loewi, 2010] that social capital is far better understood as a phenomenon of information access, and believe that McFarland and Thomas' findings support this. In this paper, political engagement seems not to be a function of a single activity, but rather intense engagement in a broad variety of activities, or broad knowledge base. This may also be intuitively appealing if civil society is considered as inherently "big-picture".

2 Decentralized Search

For the literature on decentralized search we first read Jon Kleinberg's early paper, *The Small World Phenomenon: An Algorithmic Perspective*[Kleinberg, 2000]. He begins by pointing out that the original small world experiments of Milgram not only imply certain properties of the network, but also the existence of certain decentralized search algorithms. That is, we both know that short paths exists and that people were able to find such paths, using some type of local information. On the assumption that individuals only possess knowledge about the locations of their neighbors, Kleinberg shows that the standard Watts and Strogatz network model [Watts and Strogatz, 1998] does not admit decentralized search algorithms of the kind experimentally observed. He then modifies the Watts and Strogatz model by adding a parameter which controls the geographic distribution of random edges (as opposed to the lattice edges). Finally, he shows that there is a particular value of this parameter which creates networks optimal for decentralized search.

We see several issues with Kleinberg's approach. For one, it is unclear whether Kleinberg is trying to find optimal decentralized algorithms for things like peer to peer networks or are whether he is trying to accurately model the Milgram experiments. In the case of the latter, it is not clear whether the rigor with which Kleinberg treats the subject is useful, as it all stands on a rather conservative estimate of local information–namely the geographic location of their neighbors alone. For example, he ignores the possibility that a node know its neighbor's proximity to the target in non-geographic ways–such as cultural or professional affiliation, which we might safely assume plays some role in the forwarding decisions in the Milgram experiment.

A separate issue related to the formalisms employed by Kleinberg concerns the role of asymptotic analysis. His claims about the insufficiency of the Watts and Strogatz model, as well as the optimality of his own, both rest upon arguments of asymptotic analysis. This is problematic because if we wish to come up with models that mimic the Milgram experiment, we have no reason to hold such models to asymptotic bounds. Instead, we ought to be interested in demonstrations of search on networks of similar size and under similar condi-

tions. While some network metrics are invariant with respect to the number of nodes, it is not obvious that a search algorithm should be held to these standards. In other words, without rerunning the Milgram experiment on networks of different sizes, we cannot know how the real world search algorithm scales. Thus, algorithmic analysis is not an appropriate way to evaluate such an search procedure, as constant factors may be excluded.

A more realistic approach to the question of decentralized search amongst people is taken by zgr Simsek and David Jensen. Their approach is distinguished by the use of homophily and neighbor degree as the primary features for message forwarding [Simsek and Jensen, 2005, Simsek and Jensen, 2008]. Unlike the Kleinberg paper, the nodes have no global geographic mapping. Instead, the concept of homophily serves as a more general framework in which geographic clustering can be represented. Homophily is a property of a network in which two nodes that have similarity (in some non-network dimension) are more likely to be connected. It is well know that real social networks exhibit strong homophily, and it is also reasonable to expect that an individual node has access to the similarity between its neighbor and an arbitrary target.

Simsek and Jensen then describe an algorithm they call Expectation Value Navigation (EVN), in which a message is forwarded to the local contact who possesses the lowest expected distance to the target. They model the distribution over distances to the target as a binomial distribution in which a node with degree d has d independent trials with probability $q_s t$ (homophily between node s and t) to reach the target node t. Moreover, as opposed to calculating the complete recurrence, they take advantage of the fact that the probability of reaching the target node in i steps is highly correlated with the probability of reaching the target in $i-1$ steps. Thus, they propose using only the probability of reaching the target in 1 step as their decision criterion, which yields a simple algorithm that combines only local knowledge of homophily and neighbor degree at each step. The results show that it is significantly better than an algorithms based upon either degree or homophily alone and that it approaches the performance of global depth first search under certain circumstances.

One critique of the Simsek and Jensen paper is that they construct their test graphs using the same parameter which is used in navigating the network. It seems like this gives the algorithm something of a head start and might be best to control for. Another way to put this is that in order for this to be a realistic model of social navigation, people must be able to judge similarity between their neighbors and a target in the very dimensions which gives rise to similarity-based edges in the network structure.

[Adamic and Adar, 2005] report on research done in the usability of network abstractions for a relatively small real-world network. In the paper the authors combine information about email communication between employees of HP Labs with both spatial information about their workplaces and information about the hierarchy of the office. They verify that the data set exhibits the small world structure and that it is searchable using decentralized search algorithms (as defined in [Watts et al., 2002]). They also evaluate different search strategies to do so, each using different combination of the available networks. This approach

of combining multiple information layers can be very interesting for a range of different applications like data mining, statistical analysis, link prediction and decentralized search. However, these possible applications of mapping different networks together is not further explored in the paper. They also use a relatively limited set of three different networks. Both of these points can be improved upon.

3 Proposal

Within the group there is interest concentrated around both the social ramifications of networks, and exploring the possible mechanisms of search within a network. We are attempting to treat these two topics simultaneously, and believe this will be entirely possible, but have not yet decided precisely how the topics will be merged. A presently discussed possibility is to use the metadata to determine distances between students, and then run a variety of search algorithms on the network, in an attempt to recreate the interaction data. Both the distances (varying the relative importance of various parameters) and the algorithm can be varied in an attempt to find the best approximation, which if feasible would lend credence to the notion that the particular search algorithm was being used within the social space. In general, we are interested in casting some social behavior as a search problem, which we will then attempt to model in simulations.

Alternatively, it would be interesting to try constructing networks possessing homophily with respect to certain node attributes between nodes and then to attempt to search those networks using some correlated attribute, but not the attribute by which the synthetic network was generated in the first place. This would simulate the way in which our knowledge of the similarity between a neighbor and the target is usually only approximate and may not correspond to the same concept which is operative in whatever dynamic actually causes the network to exhibit homophily.

Another area of potential application is in the testing of hypotheses formed in response to McFarland and Thomas' work. As we have both interactions and club affiliations in the data set of interactions between high school students with which we will be working, it will be possible to test the correlation of network variety and indicators of future civic engagement. Such a finding could lead to very deep-reaching policy recommendations regarding the importance of diversity and intensity of mandatory curriculum and other opportunities, rather than being limited to supporting a particular subset of groups that may only be coincidentally important.

References

[Adamic and Adar, 2005] Adamic, L. and Adar, E. (2005). How to search a social network. *Social Networks*, 27(3):187–203.

[Kleinberg, 2000] Kleinberg, J. (2000). The small-world phenomenon: an algorithm perspective. *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, page 170.

[Loewi, 2010] Loewi, A. (2010). The unique importance and effects of personal interactions on the general exchange of information. unpublished.

[McFarland and Thomas, 2006] McFarland, D. and Thomas, R. (2006). Bowling young: How youth voluntary associations influence adult political participation. *American Sociological Review*, 71(3):401–425.

[Putnam, 2000] Putnam, R. D. (2000). *Bowling alone: the collapse and revival of American community*.

[Simsek and Jensen, 2005] Simsek, O. and Jensen, D. (2005). Decentralized search in networks using homophily and degree disparity. *International Joint Conference on Artificial Intelligence*, 19:304.

[Simsek and Jensen, 2008] Simsek, O. and Jensen, D. (2008). Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35):12758.

[Watts et al., 2002] Watts, D., Dodds, P., and Newman, M. (2002). Identity and search in social networks. *Science*.

[Watts and Strogatz, 1998] Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442.

Very well written report. Extensive analysis, and good research with possible future applications of your idea.

Score - 100 / 100.