Presented at the AAAI Spring Symposium - Human Behavior Modeling, March 2009, Palo Alto CA

Modeling Social Diffusion Phenomena using Reality Mining

Anmol Madan and Alex (Sandy) Pentland

MIT Media Laboratory
20 Ames St. Cambridge MA 02139
{anmol, sandy} @ media.mit.edu

Abstract

The *diffusion* of information, ideas, opinions and media in a social network and the *influence* of individual nodes on the diffusion process are important questions in the social sciences. However, till date, there has been no method to automatically capture fine-grained social interactions between people and utilize it to better model the diffusion process. In this paper, we describe the use of socially-aware mobile phones to capture face-to-face interactions and music diffusion for eighty-percent of the residents of an undergraduate dormitory. We show that observations of diffusion can be used as an `active probe' to parse the network structure and the type of relationship between nodes more accurately than with 'passive' mobile sensor data alone. We propose that automatically captured social interactions can be used to create more accurate quantitative models of real-world diffusion and influence.

Introduction

Social networks play a fundamental role in the propagation of ideas, opinions, innovations, recommendations and media. *Diffusion* is the phenomena of propagation within a social network. *Social influence* is the ability of a node to manipulate the propagation process, by inducing other nodes to adopt or reject the transmission.

Models of social diffusion and influence have been studied in many different forms. For example, the transmission of political opinions and news in political science [10]; the diffusion of innovations in management science [18]; the value of novel information in organizational behavior [1], the propagation of obesity and smoking behaviors in public healthcare [5]. Several simple probabilistic models of diffusion processes have been proposed, like the threshold model [13] and variants of the cascade model [11]. Social influence is also a fashionable topic of discussion in popular culture, as marketers and product designers attempt to utilize viral media and viral propagation to advance their products or services. The proliferation of social applications on the web has generated copious amounts of data about user interaction and observations of diffusion. These data, in turn, are driving new theories and a better understanding of the field [14, 19].

In order to create realistic predictive models of diffusion phenomena, it is important to train with a complete picture of the social interactions between participants and the exogenous variables that affect the transmission process. An important aspect missing from prior work is fine-grained data about communication and face-to-face interaction between individuals. Existing social science research has relied on survey instruments to capture such interaction data. However, surveys simply cannot provide fine-grained data about the user's day-to-day interactions or communication with others. In addition, human errors are induced into surveys due to time omission (i.e. the memory of events and actions decays with time), telescoping effects (i.e. individuals tend to under-estimate the time dimension of an event) and selective memory bias (e.g. people find it easier to remember social interactions with people they admire or are attracted towards). In a survey of informant accuracy literature, Bernard and colleagues found that recall of social interactions in surveys is typically in the range of 50% accuracy [2]. Similarly, Brewer and Webster found that when asked to recall something as important as friends living in the same dormitory, college students failed to even mention about 20% of their friends [3].

With better tools to capture face-to-face interaction and we could answer questions like-- if we measure who talks to whom, and how often, does that represent the transmission probability between two people? Does regular co-location or frequent communication imply greater social influence? What is the role of different types of communication and interactions, e.g. the interaction in the workplace or in social milieu – do they translate into different types of social influence? Is one type of interaction more powerful than the other?

In this paper, we show that fine-grained interactions between people in a social network captured using socially-aware mobile phones can be used to understand real-world diffusion phenomena. We describe our current experimental deployment with eighty percent participation in an undergraduate dormitory and discuss analysis and results from a smaller pilot deployment. We show that observations of diffusion can be used as an `active probe' to parse the network structure and the type of relationship between nodes more accurately than with 'passive' mobile sensor data alone. We propose that the combination of automatically captured social interactions and measured diffusion can be used to create more accurate quantitative models of diffusion and influence in real-world social networks.

Background and Related Work

Mobile Phones as Social Sensors

There have been several recent projects that have used pervasive, mass-market mobile phones as active social sensors. Eagle and Pentland [7] coined the term Reality Mining, and used mobile phone Bluetooth transceivers, phone communication logs, and cellular tower identifiers to identify the social network structure, recognize social patterns in daily user activity, infer relationships, identify socially significant locations, and model organizational rhythms. Farrahi and Gatica-Perez [8] used probabilistic topic models learned from activity-related cues in the same dataset, to identify behavioral routines in an unsupervised manner. Liao and colleagues [15] used a hierarchical Markov model and particle filtering approach to infer locations related to changes in modes of transport, destinations and novel user behavior. Gonzalez et. al [12] analyzed GPS location traces for a 100,000 individuals and found that a simple spatial probability distribution could be used to characterize human mobility patterns better than random walk or Levy flight models. Onnela and colleagues [17] used phone communication logs to characterize the local and global structure of a 4.6 million node network, and found that intermediate strength ties play a key role in the diffusion of information.

Similarly, socially-aware electronic sensor badges have also been used to capture interactions and learn the structure of social networks. Choudhury and Pentland [4] designed the Sociometer, a wearable sensor package for measuring face-to-face interaction between people using an infrared (IR) transceiver, a microphone and accelerometers. Face-to-face interactions captured using the Sociometer were used to model the structure and dynamics of social networks. The Sociometric badge [20] was designed to identify human activity patterns, analyze conversational prosody features and wirelessly communicate with radio base-stations and mobile phones. Sensor data from these badges has been used in various organizational contexts to automatically predict employees' self-assessment of job satisfaction and quality of interactions.

Relevant Theories of Diffusion & Influence

What kind of social interaction data captured from mobile phones can predict diffusion? In his theory of social influence, Friedkin [9] explains that strong, cohesive ties between nodes lead to high interpersonal influence and faster diffusion. It is likely that such strong ties will be easily detected in co-location and communication patterns of users. An alternative explanation is the theory of weak ties [13], which discusses the bridging role of long-distance ties in diffusion. Due to less-frequent interaction, such weak ties are not likely to be frequently expressed in location and communication data, or they may not be recognized as ties. Hence, in our case, it is expected that such weak ties are harder to detect from co-location and communication features available on mobile phones. Features that capture email and other online interactions may help to complete the picture. A third, more recent theory [17] suggests that medium-strength ties play the most important role in the diffusion process.

Measuring Diffusion in a Social Community

Our current study consists of sixty-five undergraduate residents of a university dormitory. They represent eighty percent of the total population of the dormitory—the remaining twenty percent of students declined to participate in our study citing privacy concerns. The undergraduate dormitory is known for its pro-technology orientation and tight-knit community.

Each participant is using a Windows Mobile smartphone for a year, modified with the following capabilities:

- ▪ The phones periodically scan for Bluetooth wireless devices in proximity. Mobile phones are equipped with class 2 Bluetooth radio transceivers, which have a maximum range of 10m. It has been shown that Bluetooth and other wireless-radio based co-location techniques can be used to identify the nodes and edges in the social network [4, 7 and 19].
- The phones periodically scan for wifi (WLAN 802.11b) access point identifiers. Since the university campus has high wifi penetration, these identifiers can be used to infer homogeneity and entropy of location and proximity patterns, e.g. is there a cluster of users who tend to visit similar locations frequently? In addition to measuring co-location between participants, we also compute the statistical distance between the distributions of locations frequented by different participants.
- All phone call logs and sms logs are captured. The temporal and frequency features extracted from communication logs can be used to infer strength of social ties and identify relationships, e.g. how

often do certain people call on weekends?

- A custom music player is installed on the phone, which allows participants to play, share, rate and search through the music library. Participants have access to over 1500 independent music tracks from a wide assortment of genres. All events are logged on the server-side, and user-ratings are used to control for quality in the analysis. To send a track to any other participant, participants simply click on the 'share' button on the mobile phone application and select the recipient.

- To eliminate confounding effects, special care was taken to ensure that the artists and albums distributed through the custom music service were not featured in mass media or were otherwise familiar to the participants. All the content was sourced under the Creative Commons license or with explicit permission from the independent artists.

In addition to the music propagation, participants are required to complete monthly surveys that help us model diffusion along the following behavioral dimensions:

- Sociometric survey for relationships (choose from 'friend', 'acquaintance', or 'don't know')
- Political opinions (democratic vs. republican)
- Recent smoking behavior
- Attitudes towards exercise and fitness
- Attitudes towards diet
- Attitudes towards academic performance
- Current confidence and anxiety level

This experiment is currently in progress and we expect to have more results by the conference. Meanwhile, we report on the findings from a pilot deployment in the next section.

Pilot Deployment: Analysis and Results

Data Collection:

The mobile phone platform described above was deployed with seventeen residents of three floors of a similar undergraduate dormitory for one month. Data was captured from two sources—long-term social interaction data in the form of WLAN IDs, call and sms logs were captured using mobile phones; and the consumption and propagation of music with timestamps was logged on the music server. Data was discarded from two participants due to logging errors.

The social interaction dataset for all users over 30 days consists of 3499 unique call events (making a call, receiving a call or missed call), 350 short message (SMS) events (sending or receiving), and 570,000 snapshots of WLAN identifiers. The average length of a call is 122 seconds, 663 call events are during off-peak hours (i.e. after 11pm and before 9am), and 1154 call events are from weekends (i.e. either Saturday or Sunday). Over the entire month, 111 songs were shared and 1234 songs were played by users on mobile phones.

The following features were extracted for every participant dyad and used in the subsequent analysis of relationships and sharing behavior between 210 dyads:

- **Communication features:** total communication, off-peak communication (after 11pm and before 8am), weekend communication (Saturday and Sunday of the week), incoming versus outgoing communication and SMS communication

- **Location features:** co-location based on WLAN IDs, and the Jensen Shannon divergence between distributions of the first hundred most-frequently observed WLAN IDs between individuals.

Inferring Relationships from Phone Sensors

User self-assessments of relationships between dyads ('friend', 'acquaintance', or 'don't know') from the sociometric survey were used as training labels. The 'passive' communication and location features are correlated with the user-stated relationship ($r = 0.6$, $p<0.01$). If the number of shares is used as an 'active probe' of the social network, the correlation improves ($r=0.66$, $p < 0.01$).

The communication and location features help to be used to discriminate between different types of relationships, i.e., friends vs. acquaintances. The total communication and total number of shares between individuals are positively correlated with both friends and acquaintance types of relationships. The off-peak communication and SMS communication features were positively correlated only with the 'friend' relationships, and not with the 'acquaintance' relationships. Fig 1. illustrates the linear separability between the 'friend' and remaining relationship types.

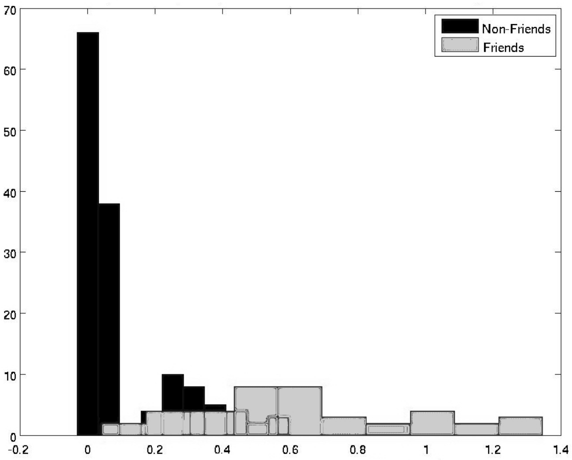

Fig 1. Histogram of 'friend' relationships vs. values predicted using the location, communication and sharing features. X-axis values are computed as a linear function of the raw features and the Y axis represents the number of dyads in each bin. Friends can be visually separated from other classes by drawing a vertical line at x=0.2 or fitting Gaussian for each class.

The classification results for predicting relationships using a BayesNet classifier and SVM classifier with 5-fold cross validation are given in Table 1. The first row is the 'non-friends' class and the second row is the 'friends' class. The training data is unbalanced (since only 28% of all dyads are friends), so a cost-sensitive approach is used in model training and classification errors for the 'friends' class were penalized more than the 'non-friends' class by a factor of 3. It is seen below that while the BayesNet classifier outperforms SVM approach in overall accuracy, recall for the 'friends' class is slightly better with the SVM model. As shown in Table 2, with the use of sharing data as an additional feature, the classification accuracy increases by a few percentage points.

Table 1. Relationship classification accuracy without using sharing data

Model	**Acc.**	**Class**	**Precision**	**Recall**	**F-Measure**
Cost-sensitive BayesNet with 5-fold CV	87.3%	Non-Friends	0.879	0.972	0.923
		Friends	0.84	0.525	0.646
Cost-sensitive Support Vector Machine with 5-fold CV (polynomial kernel)	83%	Non-Friends	0.888	0.894	0.891
		Friends	0.615	0.6	0.608

Table 2. Relationship classification accuracy after the addition of sharing data

Model	**Acc.**	**Class**	**Precision**	**Recall**	**F-Measure**
Cost-sensitive BayesNet with 5-fold CV	90.1%	Non-Friends	0.897	0.986	0.94
		Friends	0.923	0.6	0.727
Cost-sensitive Support Vector Machine with 5-fold CV (polynomial kernel)	89.6%	Non-Friends	0.942	0.923	0.932
		Friends	0.744	0.8	0.771

Estimating Likelihood of Transmission between Dyads from Social Interaction Data

Overall, the communication and location features extracted from mobile phone logs are correlated with observed sharing behaviour (r = 0.65, p < 0.01). The specific features that are important predictors of sharing are total calls and total off-peak duration, SMS communication and co-location based on WLAN identifiers. Dyadic sharing behaviour shows a higher correlation with automatically captured communication and location features than self-reported relationships (r = 0.42, p <0.01 for 'mutually acknowledged friends'). This result indicates that social interactions automatically captured using mobile phone sensors may be better predictors of the transmission probability than user self-assessments.

The media propagation observed in the experiment was further broken down into two distinct types:

(a) Approximately 70% of the total shares were between 'mutually acknowledged friends'. For this subset of dyads, the correlation of location and communication features with propagation is even higher. This represents diffusion within cohesive social ties.

(b) The remaining 30% of shares were between strangers or weak ties. For this subset of dyads, the location and communication features are not significantly correlated with sharing. This form of diffusion is consistent with the theory of weak ties.

The two types of sharing highlight the strengths and weaknesses of our approach. Face-to-face interaction features captured using socially-aware mobile phones can predict transmission probability for cohesive ties; however they are not very useful in identifying weak ties or the propagation probabilities associated with them. Other approaches like mapping email interactions or social network sites may be more useful. The Author-Recipient-Topic (ART) model and Latent Dirchlet Allocation (LDA) are examples of approaches that have been used to identify roles, relationships and group membership from email interactions [16].

The observations of sharing between participants can be broken into a 2-class (sharing /no-sharing) or 3-class model (no sharing; < 3 songs shared as 'low sharing'; >= 3 songs shared as 'high sharing'; these class boundaries were selected based on distribution of shares). Without any prior relationship data and based on mobile phone features alone, the 2-class prediction accuracy using a cost-sensitive Bayesian network classifier is 71.5 % (precision = 0.69, recall =0.426, f-measure = 0.527 for the sharing class). With a similar model, the 3-class, 5-fold CV accuracy is 69%.

It is also possible to implement a hierarchical Bayesian model, where relationships are inferred from mobile phone features, and then used as an additional feature to predict sharing. Our initial results show that with this approach, the 2-class classification accuracy for sharing increases slightly to 74%.

Modeling Social Influence

The latent-state influence model [6] is a tractable approximation for hidden Markov modelling of multiple interacting stochastic processes. In a hidden Markov model of n interacting processes, the number of latent states is product of the number of latent states per process, which implies that an impractical number of model parameters have to be learnt. In the corresponding influence model, the number of model parameters is reduced as the latent state distributions for time t + 1 are based on a linear combination of the latent states for time t. The static weights for this linear combination are the 'influence' values, and reflect the coupling between the interacting Markov chains.

$$P(s^{i}_{t+1} / s^{1}_{t}, s^{2}_{t} \ldots s^{n}_{t}) = \Sigma\, \boldsymbol{\alpha}^{ij}\, P(s^{i}_{t+1} / s^{j}_{t})$$

where:

s^{i}_{t} : hidden state of chain *i* at time *t*,

$\boldsymbol{\alpha}^{ij}$: influence of chain *i* on chain *j*, for *n* chains with *k* states per chain.

The forward backward algorithm for latent state estimation and the maximum likelihood algorithm for estimation of model parameters of the influence model are derived from the equivalence between the influence model and corresponding hidden Markov model, and the detailed derivation is available here [6].

The consumption and propagation dynamics between participants over time can be modelled as n interacting Markov chains using the influence model. Each participant represents a chain, and the observed variable is a function of captured interactions with other participants or media consumption. The inter-chain influences then represent the 'social influence' between the nodes.

Fig 2. shows the influence values for sixteen participants based on observed media consumption. The observed variable is the number of times a participant played one of the three most popular tracks. Two latent states are assumed per chain and represent the level of 'activation' for the participant. Each chain evolves with a time-step equal to one day.

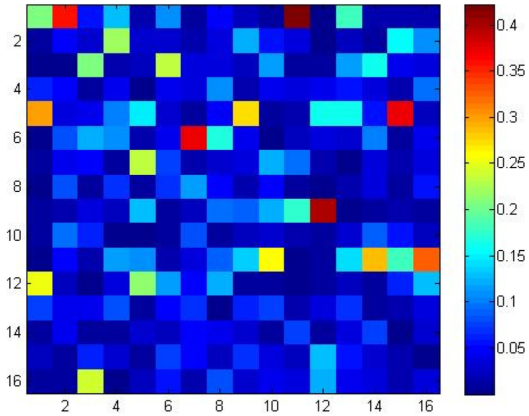

Fig 2. Social influence matrix between 16 participants -- the observed variable for each chain is the number of times the three most popular tracks are played by the participant (on a daily basis)

Self-influences are absent from this graph because the playback sequences per person for the three most-popular tracks are sparse. An interesting future direction is to explore is whether the computed influence values can be related to the observed transmission probability.

Conclusions

Our initial analysis of the use of socially aware mobile phones to predict the diffusion process is promising. We find that 70% of the sharing behavior between dyads is highly correlated with captured social interaction features, which is consistent with the theory of influence due to social cohesion between close ties. The remaining 30% of sharing behavior is between weak ties or strangers, and the social interaction features are not correlated with observed diffusion.

We also find that that the sharing behavior has a higher correlation with automatically captured social interaction features than user self-reported relationship surveys. This suggests that socially-aware systems may be better estimators of real-world diffusion and dyadic transmission probability than user's self-assessments themselves.

We show that 'passive' interaction features like phone communication logs and SMS logs are correlated with stated relationships. However, we also find that the correlations increase when we consider 'active probes' in the network, in the form of observed diffusion. The media consumption between nodes can be used to estimate the 'influences' between dyads.

Our main experiment is currently in progress and we expect to have more results by the time of the conference.

Acknowledgements

The authors would like to thank Iolanthe Chronis and Nevada Sanchez for their engineering efforts.

References

1. Aral S., Brynjolfssen E., Alstyne M.V. 2007. Productivity Effects of Information Diffusion in Networks, *MIT Center for Digital Business* , paper 234.

2. Bernard HR., Killworth P., Kronenfeld D., Sailer L., 1984. The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Reviews in Anthropology,* 1984.

3. Brewer DD., Webster CM 2000. The Forgetting of Friends and its Effects on Measuring Friendship Networks. *Social Networks*, 21, 4 pp 361-373.

4. Choudhury T., Pentland A., 2004. Characterizing Social Networks using the Sociometer. *Proc. N. Amer. Asso. Computational Social and Organizational Science*. June 2004, Pittsburg, Pennsylvania

5. Christakis N., Fowler J., 2007. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357 pp 370-379 2007

6. Dong W., Pentland A,, 2007. Modeling Influence Between Experts. *AI For Human Computing*, LNAI 4451, pp 170-189 2007

7. Eagle N., Pentland A., 2006. Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing*, Vol 10, #4, 255-26

8. Farrahi K., Gatica-Perez D., 2008. What did you do today? Discovering Daily Routines from Large-Scale Mobile Data. *Proc. ACM Int. Conf. on MultiMedia*, Vancouver Oct. 2008

9. Friedkin N.E., 1998. *A Structural Theory of Social Influence*. Cambridge University Press.

10. Huckfeldt R, Sprague J, 1991. Discussant Effects on Vote Choice: Intimacy, Structure and Interdependence. *The Journal of Politics*, 53, 1, 122-158.

11. Goldenberg J., Libai B., and Muller E., 2001. Tale of the network: A complex systems look the underlying process of word-of-mouth. *Mark. Letters* 3, 12, 211-23.

12. Gonzalez M., Hindalgo C., Barabasi A-L 2008. Understanding Human Mobility Patterns. *Nature* 453, pp 779-782 2008.

13. Granovetter M. 1978. Threshold models of collective behavior. *American Journal of Sociology*, 83, 6, 1420-1443

14. Leskovec J., Adamic L., Huberman B.A. 2007. The dynamics of viral marketing. *ACM Trans. Web.*, 1, 1, 2007

15. Liao L., Patterson D., Fox D., Kautz H., 2007. Learning and Inferring Transportation Routines. *Artificial Intelligence*, 2007.

16. McCallum A., Wang X., Corrada-Emmanuel A 2007. Topic and Role Discovery in Social Netowrks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research 30,* 249-272.

17. Onnela J.-P., Saramäki J., Hyvönen J., Szabó G., Lazer D., Kaski K., Kertész J., and Barabási A.-L.2007. Structure and Tie-strengths in Mobile Communication Networks. *PNAS 104,* 7332-7336.

18. Rogers EM, 1995. *Diffusion of Innovations.* New York, Free Press

19. Salganik M., Dodds P., Watts D., 2006. Experimental Study of Inequality and unpredictability in an Artificial Cultural Market, *Science* 311, 2006.

20. Waber B., Olguin DO., Kim T., Mohan A., Ara K., and Pentland A 2007. Organizational Engineering using Sociometric Badges. *NetSci 2007*, NYC, USA.