How "Birds of a Feather Flock Together" on Online Social Spaces

Munmun De Choudhury
School of Computing, Informatics & Decision Systems Engineering
Arizona State University
Tempe
munmun@asu.edu

ABSTRACT

Over several decades, social scientists have been interested in the idea that *similarity breeds connection*—precisely known as "homophily". Homophily structures the ego-networks of individuals and impacts their communication behavior. It is therefore likely to effect the mechanisms in which information propagates among them. To this effect, we investigate the interplay between homophily along diverse user attributes and the information diffusion process on social media. Our experiments conducted on a large Twitter dataset indicate that the particular attribute that can best explain diffusion depends upon the diffusion metric as well as the topic under consideration. In short, attribute homophily based prediction is able to quantify the actual diffusion characteristics and external trends by a significant $\sim 15-25\%$ lower distortion compared to the case when homophily is not considered.

1. INTRODUCTION

A primary domain of interest to social researchers through several decades has been the study of interpersonal communication among groups of individuals. Communication is central to the evolution of social systems. Hence the monotonic surge of interest springs from the potential of such communication impacting social processes: e.g. propagation of influence, evolution of communities etc.

Typically studies geared towards understanding these social processes via communication, until a few years back, have essentially been cross-sectional in nature, often based on participant observations [8, 2] and surveys [1] on relatively small sets of people. However, the advent of the "social web" over the past decade is providing researchers with newer ways to validate their hypotheses on large-scale data [4, 5]. For example, the Web 2.0 technologies today have provided considerable leeway to a rich rubric of platforms that promote multifaceted user interactions on shared spaces. The resultant impact of these plethora of social websites such as Flickr, YouTube, Twitter, Digg, Facebook and the Blogosphere have been widespread. Right from shopping a new car, to getting suggestions on investment, searching for the next holiday destination or even planning their next meal out, people have started to rely heavily on opinions expressed online or social resources that can provide them with useful insights into the diversely available set of options. Because electronic social data can be collected at comparatively low cost of acquisition and resource maintenance, can span over diverse populations and be acquired over extended time periods, it provides a rich and broad test bed to understand the social processes centering around interpersonal communication.

In this poster, we propose to understand online interpersonal communication process via three key aspects: "concept" i.e. the information or the "meme" that is the content of communication, the "social engagement" i.e. the social system or the network that embodies the communication process, and finally, the "channel", i.e. the media artifact via which communication takes place (Figure 1). An observed consequence of the social engagement is often the sharing of information or concepts via some media channel [3, 9, 10]. This poster deals with a specific problem of understanding, modeling and analyzing how information propagates in a social network of individuals, via observed social actions. Specifically, in this context, the poster would present the impact of the "homophily" principle on information diffusion in social media.

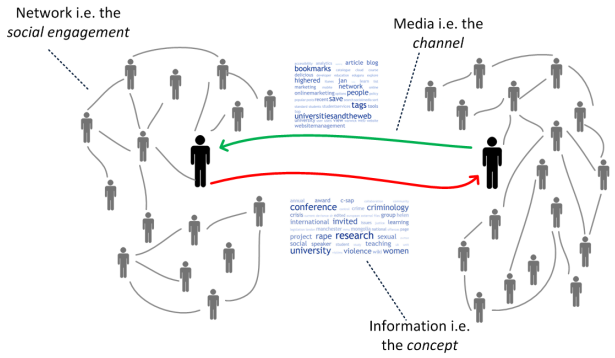

Figure 1: Illustration of the three key organizing ideas that embody online interpersonal communication processes: namely, the information or concept that is the content of communication, the social engagement i.e. the sociological framework emergent of the communication process, and the channel or the media via which communication takes place.

The homophily principle [7, 6] states that users in a so-

Submitted to Grace Hopper Conference for Women in Computing 2010.

cial system tend to bond more with ones who are "similar" to them than ones who are dissimilar. Hence homophily structures networks: people's ego-centric social networks are often homogeneous with regard to diverse social, demographic, behavioral, and intra-personal characteristics or revolve around social foci such as co-location or commonly situated activities. The existence of such homogeneity, i.e. homophily is therefore likely to impact the information these individuals receive and propagate and the communication activities they engage in. In our work, we consider communication occurring via posts on the popular micro-blogging service Twitter[1] and investigate the relationship between homophily among users and the social process of diffusion. We particularly study four kind of contextual attributes on Twitter: location, activity behavior, social role and activity distribution. Thereafter we predict diffusion characteristics under homophily on these attributes based on a novel probabilistic framework. Our experimental results on a large dataset from Twitter have been promising, and reveals how "similarity breeds connection" in a social network.

2. COMPUTATIONAL APPROACH

We propose a three step approach to investigate the role played by homophily in predicting diffusion characteristics on a given topic over time. First, we extract diffusion characteristics along different categories, such as user-based (volume, number of seeds), topology-based (reach, spread) and time (rate), corresponding to social graphs defined on different user attributes (e.g. location, activity behavior). Second, we predict the users likely to get involved in the diffusion process at a future time slice based on a Dynamic Bayesian Network based probabilistic framework. Third, we utilize the predicted set of users to determine diffusion characteristics at the future time slice. We quantitatively define distortion metrics to study how the predicted characteristics corresponding to each attribute (i.e. presence of homophily along a certain attribute) can explain the actual characteristics as well as external time-series variables—search and news trends.

3. EXPERIMENTAL STUDIES

We now present our experimental results in this section that validate the proposed framework of modeling diffusion. We utilize a dataset that is a snowball crawl from Twitter, comprising about 465K users, with 837K edges and 25.3M tweets over a time period between Oct'06 and Nov'09. For our experiments, we focus on a set of 125 randomly chosen "trending topics" that are featured on Twitter over a three month period between Sep to Nov 2009. For the ease of analysis, we organize the different trending topics into generalized themes based on the popular open source natural language processing toolkit called "OpenCalais"[2].

We discuss attribute homophily subject to variations across the different themes, and averaged over time (Oct-Nov 2009). There are two evaluation metrics of interest: (a) saturation metrics where we compare distortion between the predicted and actual diffusion characteristics at a future time slice given an attribute, and (b) utility metrics where we describe two utility measurement metrics for quantifying the relationship between the predicted diffusion characteristics on topic, and the trends of same topic obtained from external time series. Figure 2 shows that there is considerable variation in performance (in terms of saturation and utility measures) over the eight themes.

In the case of saturation measurement, we observe that the location attribute (LOC) yields high saturation measures over themes related to events that are often "local" in nature: e.g. (1) 'Sports' comprising topics such as 'NBA', 'New York Yankees', 'Chargers', 'Sehwag' and so on–each of them being of interest to users respectively from the US, NYC, San Diego and India; and (2) 'Politics' (that includes topics like 'Obama', 'Tehran' and 'Afghanistan')—all of which were associated with important, essentially local happenings during the period of our analysis. Whereas for themes that are of global importance, such as 'Social Issues' the results indicate that the attribute, information roles (IRO) yields the best performance.

From the results on utility measurement, we observe that for themes associated with current external events (e.g. 'Business-Finance', 'Politics', 'Entertainment-Culture' and 'Sports'), the attribute, activity behavior (ACT) yields high utility measures. This is because information diffusing in the network on current happenings, are often dependent upon the temporal pattern of activity of the users, i.e. their time of tweeting.

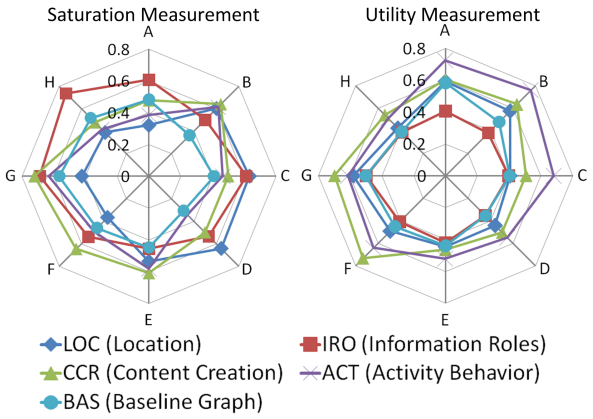

Figure 2: Mean saturation and utility measurement of predicted diffusion characteristics shown across different themes. The themes are: A–Business-Finance, B–Politics, C–Entertainment-Culture, D–Sports, E–Technology-Internet, F–Human Interest, G–Social Issues, H–Hospitality-Recreation.

From these studies, we interestingly observe that attribute homophily *indeed* impacts the diffusion process; however the particular attribute that can best explain the actual diffusion characteristics often depends upon: (1) the metric used to quantify diffusion, and the (2) topic under consideration.

4. CONCLUSIONS

We have investigated the role played by attribute homophily on the information diffusion process in online social media. To this effect, we have proposed a dynamic Bayesian network based framework to predict diffusion characteristics corresponding to different user attributes, such as location, activity behavior and information roles. Overall, attribute homophily is able to quantify the actual diffusion and external trends by a margin of $\sim 15-25\%$ lower distortion compared to cases when homophily is not considered.

[1]http://www.twitter.com/
[2]http://www.opencalais.com/

5. REFERENCES

[1] Ronald S. Burt. Structural holes and good ideas. *The American Journal of Sociology*, 110(2):349–399, 2004.

[2] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[3] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM.

[4] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[5] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[6] Miller Mcpherson, Lynn S. Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[7] Miller McPherson and Lynn Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, 52(3):370–379, 1987.

[8] Theodore Mead Newcomb. *The acquaintance process*. Holt, Rinehart and Winston, New York, NY, 1961.

[9] Thomas W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, January 1996.

[10] D. J. Watts. A simple model of global cascades on random networks. *PNAS*, 99:5766–5771, 2002.