User Generated Content Consumption and Social Networking in Knowledge-Sharing OSNs

Jake T. Lussier, Troy Raeder, and Nitesh V. Chawla

Interdisciplinary Center for Network Science and Applications
Department of Computer Science and Engineering, University of Notre Dame, USA

Abstract. Knowledge-sharing online social networks are becoming increasingly pervasive and popular. While the user-to-user interactions in these networks have received substantial attention, the consumption of user generated content has not been studied extensively. In this work, we use data gathered from `digg.com` to present novel findings and draw important sociological conclusions regarding the intimate relationship between consumption and social networking. We first demonstrate that individuals' consumption habits influence their friend networks, consistent with the concept of *homophily*. We then show that one's social network can also influence the consumption of a submission through the activation of an extended friend network. Finally, we investigate the level of reciprocity, or balance, in the network and uncover relationships that are significantly less balanced than expected.

1 Introduction

The recent emergence of online social networks (OSNs) has affected the manner in which web content is both created and used. In some cases, web sites have created entirely new *classes* of virtual content from social networks on Facebook and dynamic career profiles on LinkedIn, all the way to separate virtual worlds (Second Life). Regardless of the specific application, nearly all OSNs allow users the opportunity to create and consume content. Given the incredibly diverse range of existing OSNs, this content, commonly referred to as User Generated Content (UGC), may be the only common thread in all these networks.

As such, it is useful to characterize OSNs by the role that the UGC plays. One useful distinction, as described by Guo et al. [9], is whether an OSN is networking oriented or knowledge-sharing oriented. Networking oriented OSNs are those in which the formation and sustenance of social links are the primary concern and the sharing of UGC is only a consequence of this. Some examples of these networks are Facebook, Twitter, MySpace, and LinkedIn. In knowledge-sharing oriented networks, on the other hand, the creation and consumption of UGC is most important, and people only form social ties in order to facilitate these processes. Two examples of these networks in popular culture are Digg and Youtube. It should be noted that there is no hard line between the two classes of networks. For example, a Facebook user may occasionally friend another user simply to share information, or a blogger might friend another blogger because

of a "real-life" friendship with no intent to share information. Still, it is the primary purpose and role of an OSN which defines it.

We consider Digg, a knowledge-sharing oriented OSN. While there have been a few papers analyzing Digg content consumption [11, 19], these works deal primarily with the characterization of future consumption behavior based on past behavior, touching only tangentially on network aspects. Our work brings networking to the forefront and thus presents a more complete understanding of the relationship between UGC and social networking.

Contributions
The details about the data acquired from the social bookmarking site `digg.com` are given in Section 2. The key contributions of the paper are as follows:

1. Using a statistical measure of distributional divergence, we show evidence of *homophily* in Digg, wherein friends tend to digg stories of similar topics and non-friends' tastes are less similar. This implies that friendship on such online networks is largely a phenomenon of common interests (see Section 3.1).
2. We show that stories achieving especially high levels of consumption do so by activating the submitter's second degree friend network (see Section 3.2).
3. Using a measure of *reciprocity*, we show that UGC consumption activity can be very *imbalanced*, meaning that A consumes B's content much more readily than B consumes A's content. This implies that there are highly unbalanced dyads, leading to an important distinction between real-world human social networks versus UGC derived social networks (see Section 3.2).

2 Data Specifics

Before delving into consumption and social networking, we first present a brief depiction of `digg.com`, our particular data set, and the social network we constructed. Launched in 2004, `digg.com` was intended to democratize digital media. Digg allows users to discover and share content from anywhere on the web by pasting a URL; indicating whether it is a story, video, or image; and providing a short description. Other users then comment on the content, or simply "digg" (like) or "bury" (dislike) it. Once a submission has earned enough diggs, it becomes "popular" and jumps to the homepage in its category. Stories that are not yet popular are listed in the "upcoming" section. Finally, Digg allows users to add others to their social networks. If user A adds user B to his or her network of friends, A becomes a *fan* of B. This unidirectional link allows the initiator to monitor the other's activity. Specifically, once A selects B as a friend, A can see any stories that B submits or diggs through a special "friend" interface. If B reciprocates and returns A's friendship, then A and B are called *friends*. Since its launch, Digg has grown to over two million users and has prompted the creation and growth of other social networking sites centered on story creation and dispersion.

We performed a single crawl of digg.com, which returned 6,073,456 friend relationships and 564,193 users. We then constructed a network in which users are

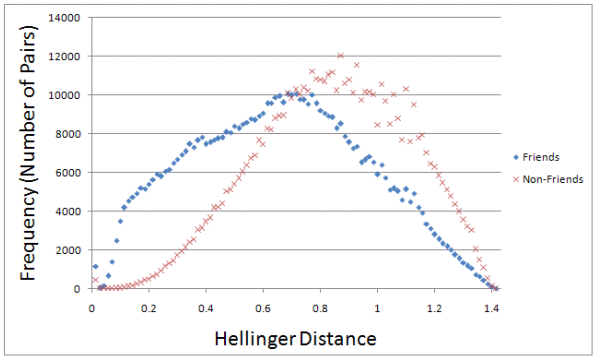

Fig. 1. Distributions of Hellinger Distances for friends and non-friends.

nodes and fan relationships are directed links. More specifically, we consider a directed edge from A to B if A has added B to his network of friends. The resulting network has 564,193 nodes, 6,073,456 edges, an average clustering coefficient of 0.075, and an average degree of 16.146. The node-degree distribution has a clear heavy tail: 78% of users have degree less than five, about 0.1% of users (567) have degree $> 1,000$ and only eight users have degree $> 10,000$. Thus, although most users have relatively few connections there are a substantial number of users who are extremely well connected.

3 Social Networks and Consumption

As stated previously, UGC consumption is a primary driving force in knowledge-sharing oriented OSNs. In this section, we will substantiate this claim by illustrating the relationships between consumption and social networking.

3.1 Consumption Patterns and Friend-Making

In order to study the relationship between social networking and consumption, we need a means of quantifying the difference between two individuals' consumption patterns. A Digg story is classified into one of several *containers* (such as Technology or World & Business) that broadly describe subject matter. If a user diggs or comments on a set of stories $\mathcal{S}$, the set $\mathcal{S}$ will form a *distribution* across the various containers. We would say that two users are similar if they comment or digg similar stories, and we can measure this similarity based on the distribution of the stories they digg.

To do this, we employ *Hellinger distance* [5], a measure of distributional divergence that is both *bounded* and *skew insensitive*, meaning that its value does not depend on the number of samples from either of the distributions being compared. The Hellinger distance $d(a, b)$ between two distributions a and b is defined as $d(a,b) = \sqrt{\sum_i \left(\sqrt{\frac{a_i}{|a|}} - \sqrt{\frac{b_i}{|b|}}\right)^2}$

where i, in this case, runs across all possible containers, a_i represents the number of diggs or comments by user a in container i, and $|a|$ is the total number of diggs or comments by user a.

Thus, users who tend to consume similar content have smaller Hellinger distances. Calculating this distance for all pairs of friends and all pairs of non-friends, we can determine whether friends tend to have similar consumption patterns. Friend and non-friend distributions of Hellinger distances are shown in Figure 1. As can be seen, the distribution of distances for friends is shifted far to the left, indicating that consumption patterns may influence friend-making behaviors. This is consistent with the sociological concept of *homophily*: that individuals tend to befriend people similar to themselves.

3.2 Controlling the Consumption of UGC

Often, the creators of user-generated content have a vested interest in the consumption of their content: people want their work to be seen. Now while Section 3.1 illustrated how consumption influences friend-making, the social-network structure within Digg, discussed in Section 2, also allows friendship networks to indirectly affect consumption. When a user submits, comments on, or diggs a story, that story becomes visible to his or her fans through a page known as the "friends interface." This gives the user's direct connections the opportunity to read the story and then comment on it or digg it. If a submitter promotes a story well and it receives enough diggs, it is "promoted" to the front page of Digg, where it is prominently displayed to casual visitors of the site.

We now study the impact of this control. Specifically, we study the importance of the submitter's friend network on the promotion of stories. Following typical Digg terminology, we will henceforth refer to stories that have been promoted as "popular" stories and those that have not been promoted as "upcoming" stories. Figure 2 plots the number of diggs in each hour of a story's lifetime for both popular and upcoming stories. It is immediately clear that appearing on the Digg front page makes a substantial difference in the consumption pattern of a story: the upcoming stories receive most of their diggs at the very beginning of their lifetime, and their consumption activity decays monotonically and rapidly (with a slight blip at 24 hours). For popular stories, digg activity increases for the first several hours then slowly decays.

For a simple explanation of this gradual increase, see Figure 3. The top graph shows comment and digg activity for popular stories relative to the times that they were promoted. We see that the time immediately after promotion is by far the busiest time for a story, with a steady, gradual decrease thereafter. The bottom graph plots the age at which stories become popular. We see a wide range of ages, with some stories achieving promotion almost instantly, while others lingered for over two days. However, most stories that achieve popularity do so very quickly: more than half of our popular stories were promoted within 10 hours of submission, and over 90% are promoted within a day. This range accounts almost exactly for the peak we see in Figure 2.

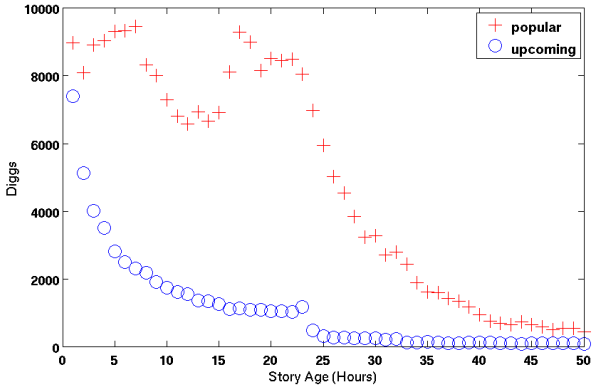

Fig. 2. Digg activity over the lifetime of popular vs. upcoming stories.

How do submitters achieve such rapid recognition for their stories? One way would be to post the Digg link on a heavily trafficked web page and rely on its visitors to digg the story. Another possibility is by means of the network mechanisms described, relying on friends and fans to spread the story. The remainder of the section studies the effects of these networks.

Distance	Pre-Promotion	Post-Promotion
0	0.0115	0.0000
1	0.4611	0.0296
2	0.2812	0.1310
3	0.1579	0.4702
>3	0.0881	0.3690

Table 1. Proportion of a story's diggs by shortest-path distance from the submitter.

The most basic evidence of a network effect on popularity is a difference in consumption patterns before and after promotion. If a person's network plays a role in the promotion of stories, we would expect that a substantial proportion of a story's diggs prior to promotion would come from members of the author's network. After promotion, by contrast, we would expect diggs fairly evenly across user population at large. Table 1 shows that this intuition plays out: over 46% of all diggs in the pre-promotion period come from direct friends of the submitter. Of additional interest is the importance of a user's second network. This group (friends-of-friends of the submitter) contributes a larger proportion of the diggs in the pre-promotion period than after promotion, suggesting that diffusion through the friend network contributes to the success of popular stories.

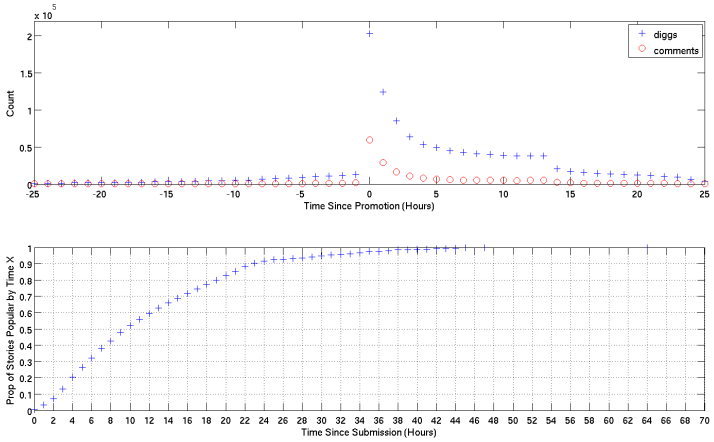

Fig. 3. [Top] Digg activity relative to promotion time. [Bottom] Promotion time.

The above result is reasonable, as the friends interface mechanism naturally supports such diffusion. Given a relationship $t \rightarrow u \rightarrow v$ between users t, u, and v, any story that t submits and u diggs will be presented to v through the friends interface. A more complete picture of this phenomenon is given in Figure 4, which shows Digg activity for popular stories as a function of the age of the story (in minutes) for several different shortest-path distances between the submitter and the person digging the story. Here, a very surprising result emerges. In popular stories, the submitter's friend network is activated *almost immediately*. While direct friends of the submitter dominate a story's digg activity for almost two hours after submission, direct-friend diggs reach their peak incredibly quickly ($\approx$ 10 minutes) and then slowly decay as the friend network becomes saturated. By contrast, we see very few diggs early on from users that are three or more steps away from the submitter. These users dominate much later in the story's lifetime as it becomes more popular and more universally accessible.

Activity among second neighbors of the submitter is elevated from very early in the story's lifetime before leveling off, providing additional evidence that the diffusion supported by the friends interface is a significant factor early on in the lifecycle of popular stories. This suggests a modification to the promotion-prediction model of [11], which casts promotion as a function of *interestingness* of the story and *number of fans* of the submitter. It may be more appropriate to consider second-neighborhood size (number of people $\leq$ two levels out) due to this diffusion effect. When friends of the submitter (who generally act quickly) digg a story it becomes visible to their friends. Some fraction of these friends will digg the story, but the effect becomes less noticeable as friends-of-friends hear about the story through other means.

While the preceeding analysis of consumption and social networking dealt with a network of friends, it is also useful to consider a network in which nodes

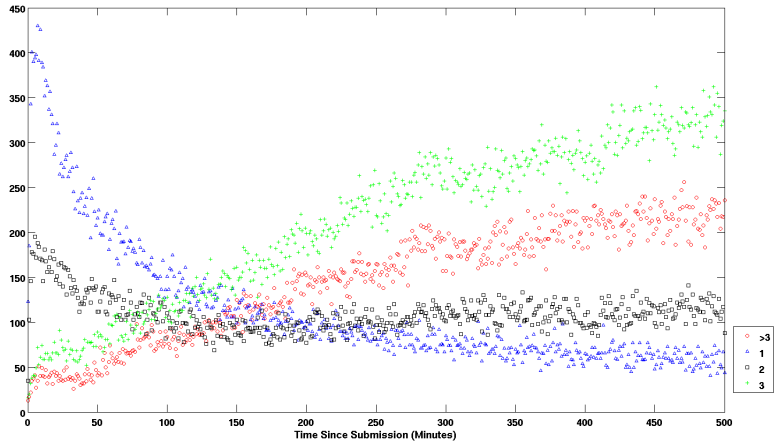

Fig. 4. Digg activity as a function of story age for different shortest-path distance.

represent users and directed edges represent an individual digging or commenting on another user's submission. Such a network allows us to consider the relationships not only between friends, but more generally between any two users who interact. We calculate a measure of *reciprocity*, or relationship balance, for any user pair for which at least one user has submitted at least five diggs or three comments. The general equation for reciprocity for any users a and b divides the number of comments or diggs from a to b by the comments or diggs from b to a. Moreover, we add one to the numerator and denominator as a smoothing factor. The actual equation is given by:

$$reciprocity(a,b) = \frac{consumptions_{a \to b} + 1}{consumptions_{b \to a} + 1} \tag{1}$$

After doing this for all eligible pairs, we then take the logarithm of all the reciprocities so as to transform them into pairs of equal but opposite values. The distribution of these values is then binned into 100 equally sized intervals, as shown in Figure 5.

For both digg and comment distributions, many relationships are either even or mildly uneven, but of particular interest is a small set of users whose relationships are extremely uneven. For example, there are 1,003 dyads (user pairs) in the data for which digg reciprocity is at least 20. That is to say, between two users Alice and Bob, Alice diggs Bob's submissions 20 times more than Bob diggs Alice's submissions. Going out further, we find 456 relationships with reciprocity at least 30 and 302 with reciprocity at least 40.

These heavily imbalanced relationships represent a critical distinction between the (quasi-)social networks developed on UGC sites and the relationships in real-world human social networks or other OSNs. Extreme imbalance, such as 40-to-1 reciprocity, is incredibly uncommon in human social networks [6] as such relationships are generally believed to be unstable (if Bob calls Alice 40

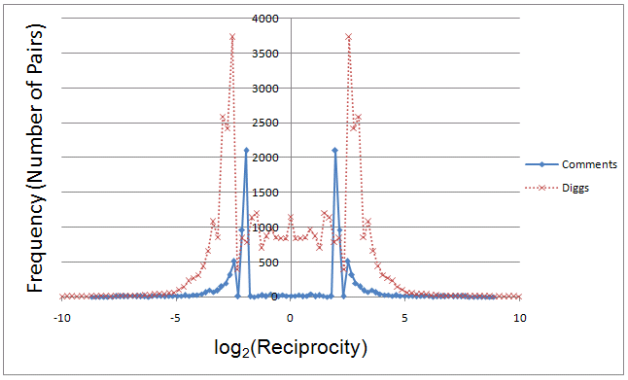

Fig. 5. Distributions of comment and digg reciprocities.

times for every time Alice calls Bob, Bob will tire of maintaining the relationship). On OSNs like Digg and iReport however, these relationships are critical to the intended function of the site, as non-reciprocity helps generate the "buzz" associated with popular articles. If all relationships on Digg were perfectly reciprocal, the only determining factor in the success of a story would be the size of the submitter's network.

4 Related Work

Much of the work done in knowledge-sharing oriented OSNs focuses on the formation [7], diffusion [2, 8, 15, 16], and growth [12, 14, 17] of social networks. Those studies that relate directly to UGC either focus on creation patterns alone or deal only superficially with consumption patterns. Cheng et al. [4] study Youtube and conclude that "related" videos have strong correlations with each other. Leskovec et al. [13] study the diffusion of news across web sites and discover that blogs generally lag mainstream news sites by only a few hours. Guo et al. [9] study UGC creation patterns and find regular temporal patterns and stretched-exponential posting behavior, suggesting that a small set of power users in knowledge-sharing oriented OSNs cannot dominate as they can in a network fitting a power-law. Agrawal et al [3] propose a method for identifying influential contributors to blogs. Certain aspects of the model (number of *in-links* a blog post receives and the number of comments it generates) are directly related to consumption. However, the authors do not study these consumption patterns directly; they merely use them as part of a larger model. Lastly, Guo et al. [10] touch on ideas related to consumption when they examine media access patterns. However, media access is simply the viewing of any form of media available on the web whether it is user generated content or not.

The studies that have examined the consumption of user-generated content focus primarily on characterizing the future consumption patterns of stories

based on past consumption. Wu and Huberman [19, 20] model the popularity of stories on `digg.com` and find that the number of diggs N_t that a story receives after time t is modeled by a simple multiplicative process. Hogg and Lerman [11] develop a stochastic modeling framework for user-generated content and use `digg.com` as an example.

5 Conclusions & Future Work

We studied *consumption* of user-generated content in OSNs in the context of the social bookmarking website `digg.com`. In contrast to other works, which have focused primarily on characterizing future consumption patterns based on past consumption, we focused more on the interplay between social network formation and content consumption.

In doing so, we showed that similar consumption patterns imply a higher likelihood of friendship. This finding provides evidence of *homophily* (the tendency of people to choose friends similar to themselves) in the Digg network.

In studying the effect of the Digg friendship network on the promotion of popular stories we have two significant findings. First, stories that are successfully promoted to the Digg front page tend to activate the submitter's friend networks very quickly, with friends of the submitter often digging a submission within minutes. Second, we find that second-level neighbors (friends-of-friends) of the submitter also figure prominently in the very early life of a story.

Finally, we studied the level of *reciprocity* or balance in Digg relationships. We found a small number of relationships with exceptionally high levels of imbalance, meaning that person A diggs person B's stories far more frequently than B diggs A's stories. We hypothesized that, while high levels of imbalance are typically unsustainable in human relationships, they are critical in OSNs because people consume much more content than they generate and a heavy-tailed popularity distribution (a small quantity of hugely popular content) is desirable.

Acknowledgments Our thanks to Kaitlin Clark, Adam Lusch, and Michael Moriarty for providing the friend data for digg.com. This work was supported in part by NSF DHB-0826958 and the Arthur J. Schmitt Foundation. Jake Lussier was also an Ateyeh Undergraduate Scholar.

References

1. L. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43. ACM New York, NY, USA, 2005.
2. E. Adar and L Adamic. Tracking information epidemics in blogspace. *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pp. 207–214, 2005.
3. N. Agrawal, H. Liu, L. Tang, P.S. Yu. Identifying Influential Bloggers in a Community. *Proceedings of WSDM 2008.*

4. X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. *Quality of Service, 2008. IWQoS 2008. 16th International Workshop*, pp. 229–238, 2008.
5. D.A. Cieslak, N.V. Chawla. Detecting fractures in classifier performance. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pp. 123–132. IEEE Computer Society Washington, DC, USA, 2007.
6. A.W. Gouldner The norm of reciprocity: A preliminary statement. *American sociological review*, pp. 161–178, 1960.
7. R. Gross and A. Acquisti. Information revelation and privacy in online social networks *ACM workshop on Privacy in the Electronic Society*, pp. 71–80. New York, NY, USA, 2005. ACM.
8. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *Proceedings of the 13th international conference on World Wide Web*, pp. 491–501. ACM New York, NY, USA, 2004.
9. L. Guo, E. Tan, S. Chen, X. Zhang, and Y.E. Zhao. Analyzing patterns of user content generation in online social networks. *Proceedings of KDD 2009*, pp. 369–378. ACM New York, NY, USA, 2009.
10. L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang. The stretched exponential distribution of internet media access patterns. *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, pp. 283–294. ACM New York, NY, USA, 2008.
11. T. Hogg and K. Lerman. Stochastic Models of User-Contributory Web Sites. *AAAI Conference on Weblogs and Social Media*, 2007.
12. R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks *Proceedings of KDD 2006*, pp. 611–617. ACM New York, NY, USA, 2006.
13. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. *Proceedings of KDD 2009*, pp. 497–506. ACM New York, NY, USA, 2009.
14. J. Leskovec, L. Backstrom, R. Kumar, and A Tomkins. Microscopic evolution of social networks. *Proceedings of KDD '08*, pp. 462–470. New York, NY, USA, 2008. ACM.
15. J. Leskovec, L.-A. Adamic, B.-A. Huberman. The Dynamics of Viral Marketing. *ACM Transactions on the Web*, 2007.
16. J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. *SIAM International Conference on Data Mining (SDM 2007)*, 2007.
17. D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, 2005.
18. J.R. Quinlan *C4. 5: programs for machine learning*. Morgan Kaufmann, 2003.
19. F. Wu and B.A. Huberman. Novelty and collective attention *Proceedings of the National Academy of Sciences*, 104(45):17599, 2007.
20. F. Wu and B.A. Huberman. Popularity, novelty and attention *ACM Conference on Electronic Commerce*, 2008.