GIS: Geographical Information Cascade in Online Social Networks

Eiko Yoneki
University of Cambridge

Network shared User Generated Content (UGC) is increasingly popular due to the emergence of Web 2.0. YouTube videos are an example, where the popularity of videos is driven by viewers and their popularity can vary dynamically and often dramatically. Access to the web containing UGC is known to follow a heavy-tailed distribution. For example, the top 10% of the videos in a normal video-on-demand system account for approximately 60% of accesses, and the rest of the videos (the 90% in the tail) account for 40%.

The emergence of large-scale multimedia social network communities (e.g. Facebook and YouTube) creates millions of users forming a dynamically changing infrastructure to share multimedia content. This proliferation of multimedia data spawns a technological revolution in the entertainment and media industries, brings new experiences to users, and introduces the new concept of web-based social networking communities.

In large-scale multimedia social networks, millions of users actively interact with each other, and such user dynamics not only influence each individual user but also affect the system performance. An example is peer-to-peer (P2P) file sharing systems, where users cooperate with each other to provide an inexpensive, scalable and robust platform for distributed data sharing. Thus, traditional content caching and distribution algorithms for web data are not efficient for UGC, since they are optimised for globally popular content. We need a new way to predict which contents will become popular so that the dynamics of popularity can be modelled.

Studying various aspects of user dynamics in multimedia social networks offers a comprehensive coverage of behaviour modelling and analysis from signal processing perspective. In this paper, we show that how possibly information can be spread among online social network communities and the information encoded in social network structure can be used to predict access patterns which may be partly driven by viral information dissemination, termed as a social cascade [1]. Specifically, knowledge about the number and location of friends of previous users is used to generate hints that enable placing replicas closer to future accesses.

In general, knowledge of UGC can spread in two ways: broadcast highlights or viral propagation (see Table 1). The rest happens when the UGC object is featured or highlighted on a central index page. Examples include being featured on the home page of the hosting sites (such as the featured videos list on YouTube) being promoted on an external social bookmarking site (e.g. if slashdotted) or ranking high on a Google search. UGC objects in this class have to be popular according to the indexing algorithm used.

The second possible means of propagation is by word-of-mouth, sharing explicitly with a group of friends. This can happen through online social networks, emails, or out-of-band (face-to-face) conversations. This kind of viral propagation has been termed a social cascade and is considered to be an important reason for UGC information dissemination. The links between friends on an online social network explicitly capture the means of propagation for social cascades. Furthermore, many social networking sites include approximate geography information. Thus, information about the friends of previous users and their geographical affiliations could be used to predict the geographical access patterns of future users.

When users access a UGC object influenced by their friends, it can be modelled as if infected by the friends' opinions. We envision that many ideas, messages, and products could be spread rapidly through our population as social epidemics. A recent example is the use of the hashtag 'susanboyle" on Twitter messages sent across the UK in June, 2009. Although there was no prior agreement on using this string, it quickly spread amongst Twitter users, and became the most popular hashtag. At its height, during 2 hours of period, nearly 2000 Twitter posts used the tag, making it the most popular hashtag.

We investigate how information can spread across geographies as an epidemic. We take an empirical approach, using friend lists from an online social network (i.e. Facebook) to emulate a social epidemic. We crawled two sub-networks from Stanford and Harvard Universities. For Harvard students, the first 35,000 Facebook profiles belong to past students who were the first users Approximately 20,000 users with 2.1 million links now exist. The mean degree is 63, and a maximum degree is 911. See the degree distribution in Figure 1. The users have 1,660 distinct affiliations, of which 1,181 could be mapped to geographic locations all over the globe. We select a single user as an initial infectious user and propagate the infection process to their friends. This process is repeated over n rounds, with infection spreading from the initial seed to nodes n hops away.

This study shows two possible geographic distributions of infected users. First, a rapidly shifting epidemic is observed (see Figure 2), where the infected population and the regional spread of the users changes from the third round to the fifth round. Second, the infection can also proceed without much change in geographic locations but rapid epidemic rate (see Figure 3). The history of past locations can trivially predict the future when the epidemic is localised.

Social cascade prediction predicts the geographic location of social cascades by utilising friendship and geographic information in social networks. Lacking accurate and complete geographic affiliation records in current online social networks, we use users' network affiliations and attach geographic locations to them. Success naturally depends on accuracy of geo-coding systems [5] - while the current crop of geo-coding APIs are very good at parsing, there are limitations. (e.g. \MIT", \BYU" etc. were parsed to latitude-longitude coordinates, but \SUNY Buffalo Graduate Center" proved to be too complex). Also, we are conflating geographical closeness between server replica and user, with good network connectivity. This may not necessarily be a correct assumption in all cases. We use the logical OR of a user's geographical affiliations on Facebook. On the one hand, this is beneficial because it captures information not in the social network about means for social cascade (e.g. a user might spread information to someone not on their Facebook profile but in their geographical affiliation region).

We conclude by emphasizing that identifying the geographical locations of potential next users is only half the problem. The other half of the problem is actually provisioning a server or servers such that the service time is minimised. This is a complex problem in itself, and this paper does not address all the details. Our next steps to a complete system will require resolving the question of which objects to replicate when replicas have limited storage and bandwidth, as well as possible strategies for the replicated videos replacing other videos at the replica site. Finally, our early prototype captures social cascades using a very simple model. Considerably more sophisticated models have been proposed [2,3,4]. Incorporating these could lead to better geographical access pattern predictions with our proposed approach.

Acknowledgement. This research is funded in part by the EU grants for the Haggle project, IST-4-027918, and the SOCIALNETS project, 217141.

[1] Cha, M., Mislove, A., Adams, B., and Gummadi, K. P. Characterizing social cascades in Flickr. ACM

Workshop on Online social networks (WOSN), 2008.

[2] Kempe, D., Kleinberg, J., and Tardos, E. Maximizing the spread of influence through a social network. Proc. of ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

[3] Leskovec, J., Adamic, L. A., and Huberman, B. A. The dynamics of viral marketing. ACM Trans. Web 1, 1 (2007).

[4] MacQueen, J. B. Some methods for classi_cation and analysis of multivariate observations. Proc. Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[5] Vincenty, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review XXIII, 176, 1975.

	Awareness	Delivery Method
Popular Content	Web Ad Broadcast	CDN
Heavy-tail Content	Social Cascade	Selective Replica Placement

Table 1: Content Awareness and Delivery Method

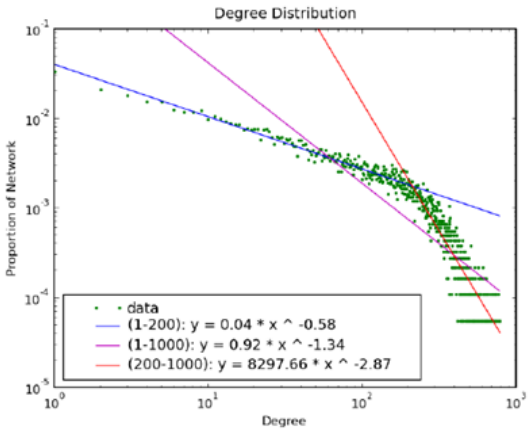

Fig 1: Degree Distribution (Harvard Data)

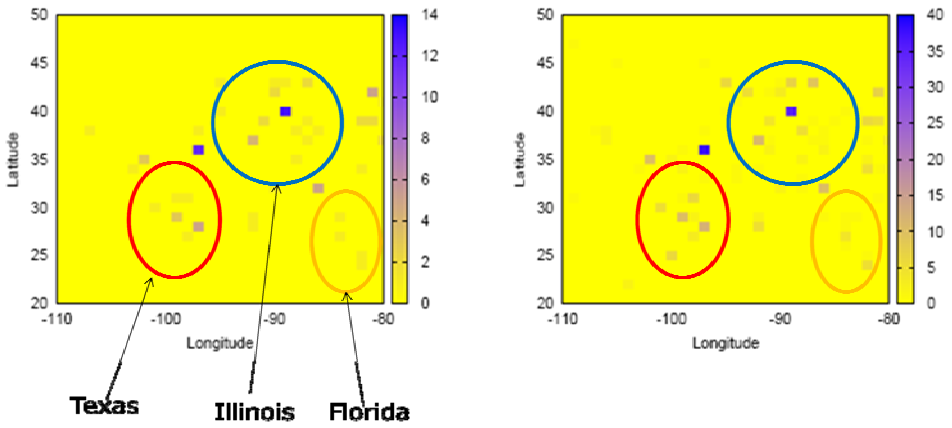

Fig 2: Shifting cascade symptom: regions of infections shifting over rounds

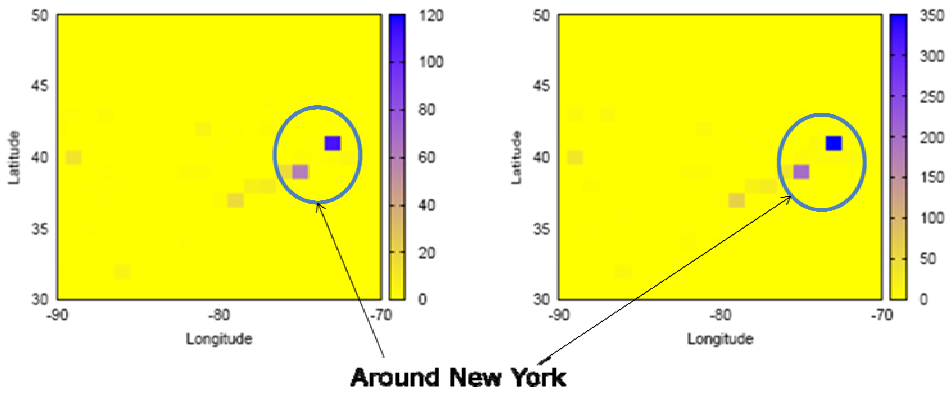

Fig 3: Stable growth of cascade: Infection regions stay the same over rounds