Topic-based Clusters in Egocentric Networks on Facebook

Lilian Weng
School of Informatics and Computing
Indiana University Bloomington

Thomas Lento
Facebook

Abstract

Homophily suggests that people tend to befriend others with shared traits, such as similar topical interests or overlapped social circles. We study how people communicate online in term of conversation topics from an egocentric viewpoint using a dataset from Facebook. We find that friends who favor similar topics form topic-based clusters; these clusters have dense connectivities, large growth rates, and little overlap.

Introduction

The principle of homophily depicts a social phenomenon that people with common characteristics are more likely to have contact with one another (McPherson, Lovin, and Cook 2001; Kossinets and Watts 2009). When combined with social influence process, this can result in a feedback loop that produces increasingly homogeneous personal networks (Crandall et al. 2008; Jamieson and Cappella 2009). Indeed, theoretical models of social influence propose that extreme polarization of interests can evolve even with a population of actors who hold diverse sets of opinions (Macy et al. 2003; Flache and Macy 2011). Meanwhile, social circles tend to restrict the range of interests that two people may discuss, further enhancing the polarization of interest groups in ego networks even more; for example, family members might be interested in all manner of life events, while coworkers might prefer business issues.

Here we scrutinize the self-reinforcing property of homophily and the polarization based on conversation topics in complex empirical environment. We observe a combined effect of homophily and intrinsic heterogeneity among social circles from an egocentric viewpoint. First, people lean towards others with preference to similar content and form dense topic-based clusters. Second, these clusters grow new internal links much faster than random chances, suggesting a feedback loop between homophily and social influence. Finally, there is no heavy overlap between clusters, implying heterogenous topical interests among friends.

Methods

We gathered a dataset with about 65,000 randomly sampled *egos* (or users) on Facebook from Apr 7, 2012 to June 16, 2012. The data collection was automatically processed on Facebook's internal servers and researchers did not try to figure out any user identity or view any text. While reconstructing an ego network, only *alters* (or friends) with at least one response to the ego's posts over the time window, denoted as *active alters*, are considered, as most egos do not actively communicate with every connected alter. Egos without any active alters are filtered out.

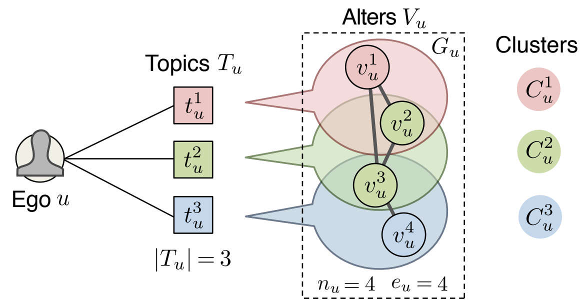

Figure 1: An illustration of definitions in an ego network and their mathematical representations. The example ego u interacts with $n_u = 4$ alters and posts about $|T_u| = 3$ topics.

Definitions

Let us first clarify several concepts in the context of ego networks. Figure 1 summarizes definitions introduced below.

Definition 1. Ego network: An *ego network* G_u is a local friendship network centered at the ego actor u. G_u contains n_u active alters of u, notated as $V_u = \{v_u^i \mid 1 \leq i \leq n_u\}$. We draw a link between v_u^i and v_u^j ($v_u^i, v_u^j \in V_u$ and $i \neq j$), if they are friends. There are e_u edges in total among n_u alters. Edges formed during the observation window are considered as *new links* and Δe_u labels the number of new links in G_u ($\Delta e_u \leq e_u$).

Definition 2. Topic: The ego u creates posts on Facebook, each post assigned with a *topic* label $t_u^k \in T_u$, where T_u includes all the topics that u has posted about.

Definition 3. Response: An alter $v_u^i \in V_u$ is deemed to be interested in topic t_u^k if she likes, comments, or shares posts about t_u^k. The intensity of v_u^i responding to t_u^k is quantified by the sum of corresponding likes, comments, and shares, labeled as $r_u^{i,k}$.

Definition 4. Topic Cluster: A *topic cluster* C_u^k stands for a sub-graph of G_u, formed by alters with responses to topic t_u^k, $\{v_u^i \mid r_u^{i,k} > 0, 1 \leq i \leq n_u\}$. The cluster C_u^k contains n_u^k nodes and e_u^k edges among which Δe_u^k are newly created over our observation window ($n_u^k \leq n_u$, $\Delta e_u^k \leq e_u^k \leq e_u$).

Topic Classification

Many Facebook posts are either not associated with descriptive text or written in short and informal wording. We therefore focus on external URL shares, as every URL directs to a Web page, providing a much richer context. Interestingly, in many cases, it is practically applicable to recognize the topic merely by checking the URL domain, seeing that pages in the same host are inclined to serve for similar purpose; for instance, *espn.com* delivers sports news and *techcrunch.com* reports on Internet products and tech companies. Top 400 most popular domains own about 85% of shares, while the most dominant domain, *youtube.com*, takes over 70% of the total shares. We thus manually assign each top domain with one of twenty predefined topic labels (see the first column in Table 1). Since Youtube videos cover an assortment of themes, we harvest Youtube category tags for each video through Google public API[1] to have a fine-grained description of video shares.

Results

We investigate three aspects of topic-based clusters in ego networks, namely dense connectivity, fast growth, and segregation of social circles.

Cluster Density: Homogeneity

The graph *density* of a topic cluster C_u^k is measured by $D(C_u^k) = \frac{2e_u^k}{n_u^k(n_u^k-1)}$.[2] Since we cannot have a direct, fair comparison between graphs with different sizes (van Wijk, Stam, and Daffertshofer 2010), we set up two baselines to simulate the formation of a cluster given the size, such that the density of a simulated module is affected by the numbers of nodes and edges in G_u but not by topics. Given a topic cluster C_u^k with n_u^k nodes and e_u^k edges, two baselines are computed as follows:

Random sampling We sample n_u^k alters at random from V_u and count the number of edges, $e_1(n_u^k)$, among them. The density of the sampled sub-graph is $D_1(C_u^k) = \frac{2e_1(n_u^k)}{n_u^k(n_u^k-1)}$. Note that we do not set the density of the whole ego network $D(G_u)$ as a baseline, because we often have $e_u^k < e_u$ and direct comparisons of graph measures between networks with different sizes can yield spurious results (van Wijk, Stam, and Daffertshofer 2010).

Weighted sampling n_u^k nodes are selected from V_n, each with probability proportional to how active the alter is. It simulates the process whereby an alter who has responded to many posts is more likely to participate in the discussion of a given topic. This baseline cluster has $e_2(n_u^k)$ edges and density $D_2(C_u^k) = \frac{2e_2(n_u^k)}{n_u^k(n_u^k-1)}$.

By comparing densities of real and simulated groups, we can estimate how concentratedly social links are allocated inside topic clusters.

Table 1 lists the average densities of empirical topic clusters and simulated ones. Irrespective of topics (see category "All" in Table 1), the average density in the data is significantly higher than both baselines, supporting the existence of topic-based homophily: alters who are interested in the same topic tend to have more contacts. The results are consistent in most single topics, except for "politics" and "religion" in which weighted sampling results in slightly higher values, though not significant. Compared with casual topics like "comedy", "knowledge", and "shopping" with evident larger densities than baselines, "politics" and "religion" are more formal and less popular. This can prevent unfamiliar alters from building up new links to reinforce the internal connectivity of the corresponding clusters; a new friendship might be easier to be initiated by common interests in movies, computer games, and designer handbags than by same partisanships and religions.

Table 1: Densities of clusters on various topics averaging across all the sampled egos. We adopt Mann-Whitney U test (Mann and Whitney 1947) to check whether the differences between the empirical measure and two baselines are statistically significant. $\langle D \rangle_u$ is compared with both $\langle D_1 \rangle_u$ and $\langle D_2 \rangle_u$, and the bigger p-value is displayed. The largest average density on each row is in bold.

Topic t_k	# Clusters	$\langle n_u^k \rangle$	$\langle D \rangle_u$	$\langle D_1 \rangle_u$	$\langle D_2 \rangle_u$	U test
All	79433	6.33	**0.331**	0.305	0.312	***
Business	782	5.37	**0.227**	0.216	0.215	*
Comedy	9473	5.60	**0.444**	0.414	0.424	***
Entertain	10918	5.43	**0.280**	0.265	0.268	**
Film	1296	4.22	**0.305**	0.278	0.282	*
Games	816	4.72	**0.246**	0.231	0.234	*
Knowledge	2645	5.63	**0.345**	0.295	0.319	***
Lifestyle	1900	4.52	**0.298**	0.261	0.267	**
Memes	308	3.57	**0.377**	0.354	0.366	
Music	8197	6.05	**0.295**	0.277	0.283	**
News	16862	6.53	**0.224**	0.215	0.220	*
Nonprofit	700	5.45	**0.259**	0.223	0.225	*
Pictures	6424	5.25	**0.512**	0.443	0.454	***
Politics	330	6.42	0.173	0.152	**0.180**	
Religion	1137	3.99	0.144	0.145	**0.147**	
Shopping	2212	4.85	**0.382**	0.329	0.335	***
Social	5251	5.83	**0.421**	0.372	0.381	***
Social tools	5940	14.44	**0.353**	0.345	0.351	
Sports	1212	5.78	**0.278**	0.251	0.261	*
Tech	626	5.04	**0.219**	0.183	0.211	*
Tools	2391	4.51	**0.490**	0.416	0.433	***

Mann-Whitney U test: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Cluster Growth: Echo Chamber

The *growth rate* of a topic cluster can be gauged by the fraction of new links established within this cluster among all the

[1] http://gdata.youtube.com/feeds/api/videos/

[2] The definition is similar to the *clustering coefficient* of the ego u in the Facebook friendship network, but the measurement of $D(C_u^k)$ only involves a portion of u's neighbors; they are equivalent when C_u^k contains all the friends of u.

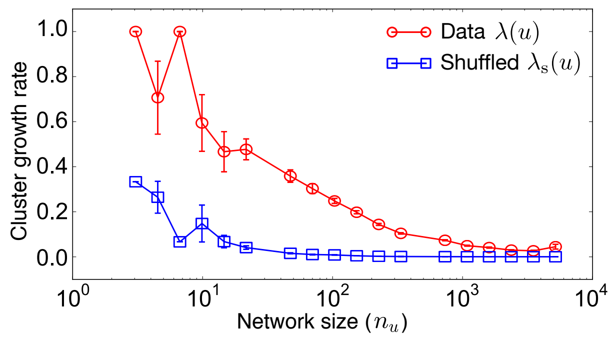

Figure 2: The plot of the cluster growth rate measured for the empirical data $\lambda(u)$ and shuffled edges $\lambda_s(u)$ as a function of the ego network size n_u.

newly formed links in the ego network. Given an ego u, we compute the cluster growth rate as $\lambda(u) = \frac{\sum_{k=1}^{|T_u|} \Delta e_u^k}{\Delta e_u}$. In the ego network observed at the beginning of the observation, there are many "missing slots" (pairs of nodes without an edge in-between) where new edges could be potentially constructed later. If the positions of Δe_u new links are random, we can simply select Δe_u missing slots at random in this early network and consider them as *shuffled* new edges. The growth rate can be similarly measured for shuffled edges, denoted as λ_s, approximating the expected cluster growth rate without topic-based homophily. We estimate[3] λ_s by:

$$\lambda_s(u) = \frac{\sum_{k=1}^{|T_u|} \left(\frac{1}{2} n_u^k (n_u^k - 1) - e_u^k + \Delta e_u^k\right)}{\frac{1}{2} n_u (n_u - 1) - e_u + \Delta e_u}.$$

where $\frac{1}{2} n_u(n_u - 1)$ is the maximum number of undirected links that can be built among n_u nodes, and the denominator $\frac{1}{2} n_u(n_u - 1) - e_u + \Delta e_u$ counts the number of missing slots at the beginning of the observation window. Similar estimation can be done for each topic cluster, as summed up in the numerator.

On average, 10.79% new edges are formed within topic clusters across all the egos, but among shuffled edges only 0.26% fall inside clusters. The high cluster growth rate in the data is robust irrespective of different ego network sizes (see Fig. 2). In conclusion, topic clusters are much more likely to gain new connections and grow dense than by chance. The shared interest between two alters in the same topic cluster largely enhances the probability of them getting to know each other later, suggesting a feedback loop between homophily and social influence (Crandall et al. 2008) or the echo chamber effect (Jamieson and Cappella 2009): frequent interactions increase similarities among group members, leading to more linkages concentrated inside the group.

Cluster Overlap: Heterogeneity

To examine whether various topic clusters are heavily overlapped, we measure the Jaccard similarity between nodes of every pair of clusters in an ego network. Given two topic clusters, C_u^k and C_u^l, the overlap is:

$$J(C_u^k, C_u^l) = \frac{|\{v_u^i \mid v_u^i \in C_u^k \wedge v_u^i \in C_u^l, 1 \le i \le n_u\}|}{|\{v_u^i \mid v_u^i \in C_u^k \vee v_u^i \in C_u^l, 1 \le i \le n_u\}|}$$

Two baselines introduced in the previous section, random and weighted sampling, can be similarly applied here. The Jaccard similarity computed for a pair of simulated groups is labeled as J_1 (random sampling) or J_2 (weighted sampling), estimating the overlap between two clusters if they are constructed without topic-based homophily. The average cluster overlap in the data, $\langle J \rangle_u = 0.0838$, is significantly smaller than what we obtain from baselines, $\langle J_1 \rangle_u = 0.1135$ and $\langle J_2 \rangle_u = 0.1316$ (Mann-Whitney U test, $p \ll 0.001$). This proposes an interesting statement that homophily causes not only homogeneity within topic clusters but also heterogeneity across clusters and topic-based segmentation: friends communicate with the ego in diverse ways owing to various individual interests and social circles, and hence they are partitioned into groups with little overlap.

Related Work

Most existing research into homophily investigated dyadic measures, pre-defined social groups, or global patterns in the network (Crandall et al. 2008; Backstrom et al. 2006; Schifanella et al. 2010). While these approaches provided a great deal of information about homophily, influence, and similarity, they did not provide an analysis from the perspective of a single actor. Here we adopt the egocentric approach, in which the local network of a central *ego* is expressed as a set of *alter* nodes connected to the ego, together with all the links among alters (Wasserman and Faust 1994). It incorporates a full set of ego's acquaintances, allowing us to estimate to what extent the friend set is segmented into distinct groups.

According to *homophily*, similar people are more likely to have contact than dissimilar ones (McPherson, Lovin, and Cook 2001; Kossinets and Watts 2009). A feedback loop is often suggested in this process to prompt the increase of similarity—people grow to resemble their friends because of the social influence and meanwhile they tend to link to similar others (Crandall et al. 2008; Jamieson and Cappella 2009). The existence of homophily in the online setting was verified through various empirical studies (Fiore and Donath 2005; Aral, Muchnik, and Sundararajan 2009; De Choudhury 2011). In the meantime, it is noteworthy that dissimilarity, disagreement, and heterogeneity also exist among people close to each other, yielding division of social groups (Brzozowski, Hogg, and Szabo 2008; Munson and Resnick 2010).

Similarity among people can be quantified in terms of various innate features (McPherson, Lovin, and Cook 2001), such as demographic characteristics, geographic locations, or *topical interests* expressed during interpersonal conversation, as employed here. Topics and user interests haven been studied broadly in the online environment from multiple perspectives (Michelson and Macskassy 2010; Romero, Tan, and Ugander 2013; Weng and Menczer 2014). *Top-*

[3]Considering that there might be overlaps between clusters, the formula slightly overestimates the growth rate for shuffled edges.

ical locality in the Web describes a phenomenon similar to homophily: most Web pages tend to link with related content (Davison 2000; Menczer 2004). Collaborative filtering (Goldberg et al. 1992), as one of the most widely adopted techniques in recommender systems, was inspired by the homophily-related observation that people with similar tastes in many items (topics) tend to have similar opinions on others. The collaborative tagging behavior of the crowd in social bookmarking applications builds up the social network autonomously on common interests in similar resources or tags. Homophily was shown to exist among users who are close in such a network while user interests and similarities are measured by the usage of vocabulary for describing objects (Schifanella et al. 2010; Aiello et al. 2012). In the measurement of social influence, existing research showed that both the topical similarity and the link structure among users should be taken into consideration (Weng et al. 2010).

Different from previous studies, we focus on the relationship between topical interests and the structure of local neighborhood. We particularly explore the ecosystem in the viewpoint of a single player, hence aiming for a better user-oriented experience in applications on social network sites.

Conclusion

We study the relationship between the friendship network structure and topical interests within a local context centered at an individual actor. Topical interests are identified by the content to which a given alter has responded. People with similar preferences are grouped, forming topic-based clusters. We show that such clusters have dense connectivities, large growth rates, and little overlap with each other. Our findings suggest that homophily reinforces the homogeneity among people with shared interests and heterogeneity across interest groups. The results are robust in complex empirical environment where individuals hold a variety of interests and form friendships under very different circumstances. We expect to further extend the method of forming topic-based clusters as well as related measures to analyze conversation topics in egocentric networks as the future work.

Acknowledgement The project was started when the first author was interning in Facebook Data Science team. We thank Facebook Data Science team for the help in collecting data and setting up experiments. We are also grateful for valuable suggestions from Prof. Filippo Menczer and NaN research group at Indiana University Bloomington. This project was supported in part by the James S. McDonnell Foundation and NSF Grant No. 1101743.

References

Aiello, L. M.; Barrat, A.; Schifanella, R.; Cattuto, C.; Markines, B.; and Menczer, F. 2012. Friendship prediction and homophily in social media. *ACM Trans. Web*.

Aral, S.; Muchnik, L.; and Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Nat. Acad. Sci.* 106(51):21544–21549.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, 44–54.

Brzozowski, M. J.; Hogg, T.; and Szabo, G. 2008. Friends and foes: ideological social networking. In *Proc. ACM Intl. Conf. on Human factors in computing systems (CHI)*, 817–820.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, 160–168.

Davison, B. D. 2000. Topical locality in the web. In *Proc. ACM SIGIR Intl. Conf. on Information retrieval (SIGIR)*, 272–279.

De Choudhury, M. 2011. Tie formation on twitter: Homophily and structure of egocentric networks. In *Proc. IEEE Intl. Conf. on Social computing (SocialCom)*, 465–470.

Fiore, A. T., and Donath, J. S. 2005. Homophily in online dating: when do you like someone like yourself? In *Proc. ACM CHI Extended Abstracts*, 1371–1374.

Flache, A., and Macy, M. W. 2011. Small worlds and cultural polarization. *J. Math. Sociol.* 35(1-3):146–176.

Goldberg, D.; Nichols, D.; Oki, B.; and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Comm. ACM* 35:61–70.

Jamieson, K. H., and Cappella, J. N. 2009. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, USA.

Kossinets, G., and Watts, D. J. 2009. Origins of homophily in an evolving social network1. *Am. J. Sociol.* 115(2):405–450.

Macy, M. W.; Kitts, J. A.; Flache, A.; and Benard, S. 2003. Polarization in dynamic networks: A hopfield model of emergent structure. *Dynamic Social Network Modeling and Analysis* 162–173.

Mann, H. B., and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals Math. Stat.* 18(1):50–60.

McPherson, M.; Lovin, L.; and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* 27(1):415–444.

Menczer, F. 2004. Lexical and semantic clustering by web links. *J. Am. Soc. Information Sci. and Technology* 55(14):1261–1269.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: a first look. In *Proc. ACM Workshop on Analytics for noisy unstructured text data*, 73–80.

Munson, S. A., and Resnick, P. 2010. Presenting diverse political opinions: how and how much. In *Proc. ACM Intl. Conf. on Human factors in computing systems (CHI)*, 1457–1466.

Romero, D. M.; Tan, C.; and Ugander, J. 2013. On the interplay between social and topical structure. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*.

Schifanella, R.; Barrat, A.; Cattuto, C.; Markines, B.; and Menczer, F. 2010. Folks in folksonomies: social link prediction from shared metadata. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, 271–280.

van Wijk, B. C. M.; Stam, C. J.; and Daffertshofer, A. 2010. Comparing brain networks of different size and connectivity density using graph theory. *PLOS ONE* 5(10):e13701.

Wasserman, S., and Faust, K. 1994. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press.

Weng, L., and Menczer, F. 2014. Topicality and social impact: Diverse messages but focused messengers. arXiv 1402.5443.

Weng, J.; Lim, E.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*.