

Linking in Social Media Does Not A Community Make

Sofus A. Macskassy
Fetch Technologies
841 Apollo St., Ste. 400
El Segundo, CA 90245
sofmac@fetch.com

Matthew Michelson
Fetch Technologies
841 Apollo St., Ste. 400
El Segundo, CA 90245
mmichelson@fetch.com

1. INTRODUCTION

Community detection algorithms have received significant attention in recent years (see, e.g., [Gibson et al., 1998, Girvan and Newman, 2002, Clauset et al., 2004, Muff et al., 2005, Newman, 2005, White and Smyth, 2005, Brandes et al., 2008, Leskovec et al., 2008, Porter et al., 2009]). The most common approaches take a graph (such as a social network) and split it into k disjoint clusters, where each cluster supposedly represents a “community” in that graph. This kind of approach is appropriate when one can reasonably expect that there is a clear enough signal in the graph, such that the found communities are likely to represent real sub-communities. For example, in Zachary’s Karate Club [Zachary, 1977], we have personal interactions between people, and we can identify the two groups that the club eventually splits into. Voting records in Congress can (with some accuracy) split into two clusters based on party affiliation (e.g., [Porter et al., 2007]), and sports-networks (team playing against team) can be split into regions (e.g., [Girvan and Newman, 2002]). In particular, this kind of approach works well on relatively well-defined, small networks, with a single well-defined and appropriate semantic interpretation to the edges. Depending on the domain, it is also important that the networks are collected and aggregated over a small timeframe.

However, the assumptions the above methods rely on start to break down when we want to identify communities in online social media such as Facebook, LinkedIn, Twitter, Digg, the Blogosphere, Flickr, etc. In these cases, the social graph is an exceedingly large and dynamic network (thousands if not millions of links and content are created every day), where relations between people are not clearly defined, and where the notion of a community itself may not be well-defined.

One cannot naïvely apply community detection based solely on the links found in these social media and expect to generate interpretable communities. Rather, one is likely to find one, enormous connected component which clusters into a few very large clusters, and a large number of very small connected components [Leskovec et al., 2008]. While we can then repeatedly apply community detection on clusters to find smaller and smaller tightly connected sub-clusters, what one eventually finds is meaningless and not a cohesive community. Further, if one were to do the same process a day or a week later, the network has changed enough that one gets completely different clusters.

Despite the difficulties above, being able to identify and characterize online communities can be incredibly useful

across a broad array of applications. Once found, we can gain deep insight into what moves the communities and their constituents making it possible to rapidly identify community-specific problems, needs, interests, etc. Even modest improvements in solving this problem can yield significant changes in how government, national security, industry and academia can use social media.

In this paper we first define the problem more formally, and then outline possible ways to address the problem, focusing on one approach in particular. The problem of finding a community in a large social media graph revolves around two core issues:

1. Defining a “community” and
2. Extracting it from the social graph.

In this paper, we define a community as a group of users that share one or more well-defined traits (such as demographics or interests).

Based on this definition, we extract communities from the social graph, considering both the users’ posted content, and the links between them. As noted above, social media links are often noisy and difficult to analyze because the social media graph has multiple types of links to consider. For instance, users might have different types of links that should be considered separately, such as friendship, family, shared-interest, school-mate, colleague, etc. Also, some links between users are not directly observable, as when users share enough characteristics to be considered as part of the same community although they do not explicitly link to each other.

There are various ways in which one can attack this problem of a large social graph with multiple types of links and user generated content.

1. We previously explored ways to tag the links in a semantically meaningful way, and break out smaller sub-graphs containing only links of the same semantic type [Macskassy, 2010]. We have shown that this does indeed yield smaller more manageable social graphs.
2. We can create new semantically useful links based on shared user content and use just these new links, or a combination of the different types of links (see, e.g., [Macskassy, 2007]).
3. We can tag the nodes in the graph based on demographic information or interests [Michelson and Macskassy, 2010], and then extract a subgraph containing only nodes which share one or more such tags.

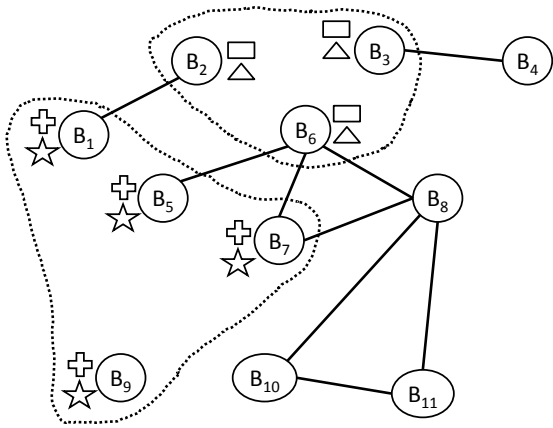


Figure 1: Graphical representation of the blogosphere. The figure shows communities defined by explicit links (lines between bloggers) which define a single large component a few tiny components. Also, the figure shows communities defined by grouping bloggers based on their demographics. There are two demographic clusters in the figure: bloggers B_2 , B_3 and B_6 with demographics triangle and square, and bloggers B_1 , B_5 , B_7 and B_9 with demographics plus and star. We contend that these demographically defined communities are more useful than those defined by explicit links.

4. We can consider methods which combine all of the above methods.

We have previously explored tagging links, creating links and combining different types of links. In this paper we explore tagging nodes with “profile” information and using these tags to extract a subgraph containing nodes which share similar graphs.

To make this notion clear, consider Figure 1. The figure shows a representative group of bloggers, each denoted as a circle and named B_i . The figure shows the explicit links between the bloggers as thick lines connecting them. The figure represents the phenomena that explicit links generally lead to one or two enormous connected components, and a set of very tiny connected components (which we verify in our experiments). As we state above, this is not necessarily the ideal definition of a community. The figure also shows some of the bloggers tagged with demographic features: notably, the bloggers B_2 , B_3 and B_6 are tagged with demographics represented by a triangle and a square, while bloggers B_1 , B_5 , B_7 and B_9 have demographic features represented by a plus and a star. We have drawn a dotted line around the communities that would be defined when considering these demographics as defining the community, in contrast to the explicit links.

As stated above, the intuition guiding this approach is that a meaningful community is one where the members have something in common. Often in a real operational setting we have some idea of what kind of commonalities we are looking for; for example, whether a group of people share common demographic attributes, have the same interests, etc. This type of information is not always di-

... Sometimes, *my kids* are geniuses...

... I basically danced a jig when *my dad* built a pen for *my dog*...

... My birthday July 10 falls on a Saturday this year, and it's *my 35th birthday* besides, ...

Figure 2: Example sentences observed from different users.

rectly observable from the network itself. However, we have a large amount of content available from all the posts provided by users. If even a fraction of this content has identifiable information, then we can start identifying well-defined communities. Combining identified profiles with the underlying linking behavior can be incredibly powerful in terms of propagating demographic information and interests.

This paper specifically explores the first question: Can we identify demographic information, and how do the demographically aligned communities differ from the explicit linking structure we see in the social media networks?

The rest of the paper is outlined as follows: We first explain our methodology for identifying and extracting demographic attributes. We then describe our methodology for identifying communities from the demographic data and our technique for comparing communities found through different approaches. We then provide a case study on blog data we have collected since January 2010. We end by discussing early work on expanding our approach to include interests as well as demographic information.

2. IDENTIFYING AND EXTRACTING DEMOGRAPHIC ATTRIBUTES

In this paper, we define a demographic as a descriptive, binary-attribute about a user. Examples might be whether a user “is-married,” or “has-pets.” For continuous or real-valued attributes, a binary demographic attribute is created by binning the real values. For instance, users’ ages can become the demographics, “Age 0-10,” “Age 11-20,” etc.

While some demographic information may be available from a user’s profile page, recent work has shown that bloggers reveal a large amount of demographic information in their own posts. For example, a blogger may refer to his or her birthday, kids, work, etc. We define such references as *self-referential facts*. Figure 2 shows a few example sentences from our data.

By aggregating the self-referential facts over all of a specific user’s posts we can generate a meaningful and precise demographic profile of that user. Therefore, our approach is to analyze millions of posts to find specific textual patterns that identify a set of well-defined demographic attributes with high precision (at the cost of recall). We emphasize high precision, as we want to ensure the fidelity of discovered communities. This approach is very useful, because as text mining improves, so will our demographic extraction and profiling capabilities.

The result of this step will therefore be a partially filled demographic profile of all users from which we have content. We provide details of these below.

3. FROM USER DEMOGRAPHICS TO COMMUNITIES

In this paper, we identify communities based on the demographic profiles we have extracted from users. Specifically, a demographic profile for each user defines an N-dimensional representation of the user’s characteristics. Therefore, we can define the similarity between users as the cosine distance between their demographic profiles. This allows us to group users together who have a high enough cosine similarity between them. Thus, the result of this step is a set of clusters of people with very similar demographic profiles.

Figure 3 shows our process for generating communities based on demographics. In the figure, we generate the demographics for three example bloggers, and form a link between two of them, defining a community based upon their similar demographic profiles.

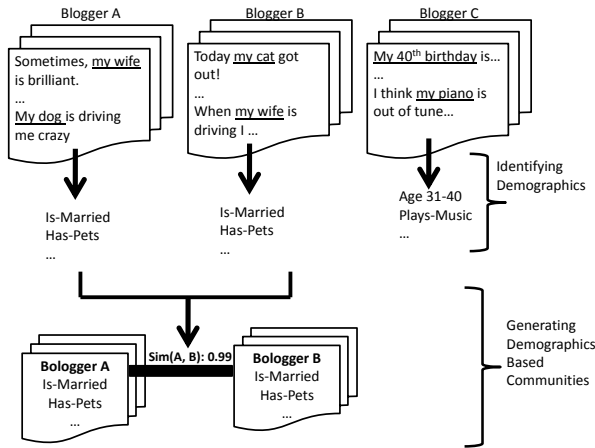


Figure 3: Demographics-based Communities

4. COMPARING COMMUNITIES

We have above outlined an approach for identifying demographically aligned communities in social media based on produced content. There are of course many other ways to identify communities, and it is often of interest to evaluate how well these different communities may align with each other.

We compute a one-sided community-alignment by taking all the communities found by one method (such as the demographic-based method above) and computing the overlap with the communities generated by another approach (such as looking at the explicit link structure).

We define the population of users as $\mathbf{U} = \{u_1, \dots, u_N\}$, and a community $c = \{u_i | u_i \in \mathbf{U}\}$ as a collection of users. We define the two sets of communities as $\mathbf{C}_A = \bigcup_i c_{a,i}$ and $\mathbf{C}_B = \bigcup_j c_{b,j}$, where $c_{x,i} \cap c_{x,j} = \emptyset$. Further, $(\mathbf{C}_A, \mathbf{C}_B) \subseteq \mathbf{U}$. In other words, each set of communities will only have a subset of the whole population.

We define the overlap of two sets of communities, \mathbf{C}_A and \mathbf{C}_B , as:

$$q(\mathbf{C}_A, \mathbf{C}_B) = \frac{1}{|\mathbf{C}'_A|} \sum_{c_i \in (\mathbf{C}'_A \subseteq \mathbf{C}_A)} \left(\sum_{u_j \in c_i} \frac{|c_i \cap \mathbf{C}_B|_{u_j}|}{|c_i|} \right), \quad (1)$$

where $\mathbf{C}_B|_{u_j}$ refers to the community in \mathbf{C}_B which u_j belongs to and \mathbf{C}'_A is the set of communities, $\mathbf{C}'_A = \bigcup_i c_i = \{u_j \in c_i | u_j \in \mathbf{C}_A \cap u_j \in \mathbf{C}_B\}$. That is, \mathbf{C}'_A only contains users that belong to both sets of communities, \mathbf{C}_A and \mathbf{C}_B .

This definition is not symmetric, and is a way to gauge how much overlap one set of communities has in another set of communities. We define overlap this way as one set of communities may have far fewer users, but each of its constituent communities could have a large overlap with a larger community in the other set. Therefore, the smaller community is well-aligned with the larger community, but not vice versa.

As an example, again consider Figure 1, which represents a blogosphere. For clarity, we will only focus on comparing the community defined by explicit links, to the demographic community defined by the features plus and star (e.g., bloggers $\{B_1, B_5, B_7, B_9\}$ in the figure). Let us first consider the overlap between the demographic community and the explicit link community. For clarity, consider bloggers B_5 and B_7 . In the explicit links communities, these bloggers belong to the community $\{B_5, B_6, B_7, B_8, B_{10}, B_{11}\}$. Therefore, the overlap for these nodes is defined as 50%, since two of the four nodes in this demographic community overlap in the same explicit link community. However, if we reverse direction, we would consider the overlap between the community $\{B_5, B_6, B_7, B_8, B_{10}, B_{11}\}$ and the demographic community, which would be 33%, since only two of the nodes from the explicit link community are found in the same demographic community. Therefore, the measure is asymmetric. We note, this computation is done for all found communities in the blogosphere. Further, the overlap between the explicit-link communities and the demographic communities are generally much smaller (as we show in our results) because the link-based communities are generally represented by a few enormous connected components.

5. BLOGGER CASE STUDY

We here explore how well the demographic communities align with the communities defined by the explicit network structure in the Blogosphere. Our hypothesis is that there will be a somewhat large overlap from the demographic communities to the link-based communities, but not vice versa. This is because we expect our demographic communities to be smaller and more cohesive than the high-variance, larger communities found by the *ad hoc* linking of bloggers. We expect the linking structure to encompass the cohesive demographic communities, but expect little overlap the other way because the link-based communities are large and not cohesive.

Our experimental data consists of blog data collected across a six month period from January 29, 2010 to July 13, 2010. The resulting blog network contains 590,011 bloggers, and 1,300,000 links, which we treat as undirected edges when creating the link-based communities.

We have content from 457,000 of the 590,000 bloggers. The other 133,000 bloggers are linked to by the bloggers we monitored and are therefore part of the blogger network even though we have not monitored them to get their content.

5.1 Demographic Communities

We generated a demographic profile for each of the 457,000 bloggers looking for specific demographic information which

Table 1: Demographics and self-referring facts

Demographic Attribute	Example phrases	Baseline
In relationship, male	“My girlfriend . . .,” “My wife . . .,”	42.18%
In relationship, female	“My boyfriend . . .,” “My husband . . .,”	57.82%
Unknown gender		95.98%
Age 0-10		9.12%
Age 11-20		17.99%
Age 21-30	“My 21st birthday . . .,”	41.75%
Age 31-40		13.33%
Age 41-50		7.72%
Age 51-60		3.86%
Age 61-70		2.81%
Age unknown		99.90%
Is Grandparent	“My grandson . . .,”	0.19%
Is Parent	“My kids . . .,”	8.86%
Is Married	“My spouse . . .,”	4.28%
Plays music	“My guitar . . .,”	0.57%
Has Pets	“My dog . . .,”	4.04%
Mentions parents	“My mum . . .,”	18.68%
Mentions grandparents	“My grandpa . . .,”	2.12%

we could extract with high precision (but also low recall). Table 1 shows the demographic attributes we extract for this case study and the baseline frequency over all users of seeing a particular demographic attribute. A user profile is then represented as a vector of binary attributes. We note that only the gender and age attributes can be missing as the other attributes are whether a particular demographic attribute was mentioned or not. For the gender and age attributes, we show the percentages of their values for when the attribute was present as well as the percentage for when it was not present. For the other attributes, a false indicates that the user did not mention whether he/she had pets, for example, but does not mean that the user does not have pets. However, we still feel that these attributes are meaningful, because if having a pet is important to a person, then chances are higher that the person would mention this in her/his blog.

To study cohesive communities based on the demographics, we studied bloggers that have at least 2 observable demographic attributes. This resulted in a subset of 36,000 bloggers from our 457,000 bloggers. We then clustered these bloggers into demographic communities by grouping them together if their cosine similarity was more than 0.99. This resulted in 212 distinct communities, ranging from 2 bloggers to 5,114 bloggers. Most of the communities found were in the range of 2-10 bloggers.

5.2 Link-based Communities

We next naïvely created a set of communities from the blogger network structure by extracting all the connected components. We removed singleton bloggers (those who had no links to other bloggers), which resulted in a set of 511,000 bloggers. From here, we extracted 15,818 connected components ranging in size from 468,476 to 2. The next largest component contained 15,817 bloggers, followed by component sizes of 292, 187, 136, and then slowly decreasing from 82 to 2. We note 98% of the clusters contained 10 or fewer bloggers. As we can see, the link structure is very much skewed to large conglomerate networks. Using community detection on those will certainly generate smaller clusters,

but they do not have as strong demographic signatures as those we found above.

5.3 Community Comparison

We next compared the two communities to better understand their interaction. Our hypothesis is that the link-structure does incorporate, to some extent, the demographic communities, but not vice versa (in large part due to the exceedingly large communities found).

We first compare the overlap from the demographic communities to that of the link-based communities ($q(\mathbf{C}_{\text{dem}}, \mathbf{C}_{\text{link}})$) and find the overlap, on average across all demographic communities, to be 0.63. In other words, for a given demographic community c_{dem} , we can find a link-based community c_{link} which contains at least 63% of the users in c_{dem} . There is a large spread of community sizes in \mathbf{C}_{dem} , and it may be that only the smaller communities were found in the large link-based community, but that the larger communities in \mathbf{C}_{dem} have much smaller overlaps. We therefore performed a micro-averaging, based on community sizes and found that the micro-averaged overlap is 0.54. This suggests that the strong demographic cohesiveness is present in the large link-structure, although it is not readily apparent. However, we can use this to expand the demographic community by taking the link structure into account, identifying other users who have a strong link-based connection to the demographic community.

To test the second part of our hypothesis, we next compare the overlap from the link-based communities to that of the demographic communities ($q(\mathbf{C}_{\text{dem}}, \mathbf{C}_{\text{link}})$). As expected, the overlap was exceedingly small (< 0.01). As with the first part of the study, we performed a micro-averaging evaluation to test whether the smaller and larger link-based communities had different overlap, but the overlap was consistently small— in fact the smaller link-based communities (< 30), 86% had zero overlap. This is a very interesting finding, suggesting that there may in fact be other stronger reasons for linking, perhaps shared interests.

In summary, our hypothesis was verified: using only explicit linking to identify communities is not possible by itself. Rather, one needs to also look at the content, and possibly the demographic profile and interest profile in order to get focused communities.

We next discuss some of the implications, and how this ought to lead to more research using content and link-structure together.

6. DISCUSSION

We have shown that links in social media cannot be used in isolation to identify meaningful communities. Rather, links must either be categorized, such that appropriate links can be used, or other measures must be used to better identify communities.

We have here investigated using the demographic profile of users. However, the content is also rich in topic-words. We have recently been exploring building a profile of users based on the things they talk about. Combining these with demographic profiling will enable us to create a very detailed profile of people who produce content, which will help us better identify niche communities.

Using the tags extracted from user content often results in sparse data on many nodes. It has been shown that some demographic information can be propagated in a blog net-

work. We can use the observed demographics to impute demographics of other bloggers. This can help us further expand a community beyond the profiles we can directly observe.

This latter part leads us back to the initial problem of identifying communities. When working on market research, politics and various types of intelligence analysis, the analyst often has a clear idea about the kind of community he/she wants to monitor. If we can build these rich profiles of users, then the analyst just needs to create a query over those profiles to get a set of users to monitor. Using the network, we can grow the set of users to include those for whom we do not have the right profile observations.

We believe marrying content analysis with network analysis is extremely powerful, and we have only begun to scratch the surface.

7. REFERENCES

- [Brandes et al., 2008] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hofer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20:172–188.
- [Clauset et al., 2004] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70. 066111.
- [Gibson et al., 1998] Gibson, D., Kleinberg, J. M., and Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (Hypertext'98)*. Expanded version at <http://www.cs.cornell.edu/home/kleinber/>.
- [Girvan and Newman, 2002] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, USA(99):7821–7826.
- [Leskovec et al., 2008] Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M. (2008). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355v1.
- [Macskassy, 2007] Macskassy, S. A. (2007). Improving learning in networked data by combining explicit and mined links. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence*.
- [Macskassy, 2010] Macskassy, S. A. (2010). Leveraging contextual information to explore posting and linking behaviors of bloggers. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.
- [Michelson and Macskassy, 2010] Michelson, M. and Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND)*, Toronto, Canada. to appear.
- [Muff et al., 2005] Muff, S., Rao, F., and Caffisch, A. (2005). Local modularity measure for network clusterizations. *Physical Review E*, 72(056107).
- [Newman, 2005] Newman, M. (2005). Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences*, pages 8577–8582.
- [Porter et al., 2007] Porter, M., Mucha, P., Newman, M., and Friend, A. (2007). Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications*, 386(1):414–438.
- [Porter et al., 2009] Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 1164–1166.
- [White and Smyth, 2005] White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.
- [Zachary, 1977] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473.