Discovering Web Communities in the Blogspace

Ying Zhou, Joseph Davis
The school of Information Technologies,
The University of Sydney
Sydney, Australia
Email: {zhouy,jdavis}@it.usyd.edu.au

Abstract

With the emergence of a range of second generation Internet based services such as weblogs and their hosting services, many loosely organized communities of bloggers have started to form around common interests. Those communities have quickly evolved into a new information and knowledge dissemination channel. Most current search engines cannot discover weblog communities through regular keyword search. In this paper we propose a new way of collecting and preparing information for weblog community discovery. Weblog community is treated as a social network and the data collection stage focuses on gaining knowledge of the strength of social ties between weblogs. The strength of social ties and the clustering feature of social network are used to extract communities from a large blogspace. We also develop a few metrics to rank communities as well as individual members in the community. We report several experimental results on "web services" communities.

1. Introduction

In recent years a wide range of second generation Internet based services such as weblogs and their hosting services emerged and gained rapid popularity among Internet users. Average Internet users are becoming more and more active in publishing on the web. Many loosely organized communities of bloggers have started to form around common interests. A prominent feature of those communities is the speed and accuracy of capturing the latest discussion and information on a range of topics from international development to specific technologies.

Internet users start to look for latest information not only from traditional new sites but also from weblogs. A case study on Technoarati, the very first and perhaps the largest blog search engine so far, reports that "by the fall of 2005, more than 15 million blogs were being read by 30 million people, and Technorati was performing millions of searches a day"[5]. In addition to instant search of information, many users also subscribe to certain weblogs and receive updates through web feeds. This is like joining a community and keeping track of what's going on in the community.

Yet it is very difficult for an outsider to discover such online communities, especially those discussing less popular topics. Currently, people get to know those communities or sites within communities either through word of mouth, or through limited categorization or listing services provided by well known websites. For instance, New York Times maintains a list of popular blogs under different categories[14]. Technorati and similar weblog search engines such as daypop.com and bloglinee.com maintain a short list of favorite blogs. Their main search services, however, only return most related pages rather than sites. Pages returned by regular search services do not necessarily mean that the whole site is about that certain topic.

One fundamental difference between regular web search and weblog search is that they are for entirely different purpose. Most of the time, web search is a one-off action to obtain information. For instance, we enter some keywords like "Voronio partition" in a search engine like Google to get the explanation or definition of this term and we may wander off after reading the related part. Only on rare occasion, we might bookmark the result page for further reference. In this case, the authorities of the pages returned are of essential importance. On the other hand, weblog search may ensure subscription of the results. For instance, by typing "web services" in a search engine like bloglines.com, the implicit purpose may be finding the community of bloggers that actively write on "web services" and keeping tracking what they are talking about for a certain period of time. Current general page ranking algorithm may not be able to identify the communities and rank them effectively.

The main contribution of this paper is the examination of weblog communities from the perspective of social network analysis. We propose a

1530-1605/07 $20.00 © 2007 IEEE

new way of collecting and preparing data for community search which emphasizes gaining informaiton on the strength of social ties between weblogs. The social tie strength data is used to as key knowledge in discovering closely knitted communities from a large set of interwoven weblogs. In particular, this paper answers the following questions:

1) How can we identify weblog communities around a particular topic?
2) How can we compare communities discussing similar topics?
3) How can we rank the members and find the important nodes in a particular community?

The rest of the paper is organized as follows. In section 2, related work on web communities and weblog search is reviewed. In section 3, we give a formal definition of blog community and its measuring metrics. Methodology regarding data collection and community extraction is described in section 4. Some experimental results are presented in section 5 with extensive discussions presented in section 6. We conclude the paper in section 7.

2. Literature Review

Web communities are defined as groups of densely connected web pages. The whole web is often modeled as a directed graph with each node representing a web page and each edge representing a link between two pages. The extraction of communities is essentially a problem of graph clustering.

The notion of authorities and hubs by Kleinberg[9] provided the basis for many pervious web community research[10][11][18]. An authority is a page with quality content on a search topic and is pointed to by many good hubs. A hub is defined as a page pointing to lots of good authorities. From this perspective, a community is viewed as a collection of interconnected hubs and authorities. An iterative algorithm HITS is given in Kleinberg's classic paper to determine a collection of webpages' hub and authority score. The PageRank[3] algorithm used in Google achieves similar goal.

Flake et al [6] gave a more general view of web communities. The web community was defined as a subset of a large connection of connected web pages that has more links within the community than links to the rest of the space. The "max flow-min cut" algorithm from graph theory was used to extract communities. The collection of web pages was obtained through a focused crawler with a few seed web pages and a predefined crawl depth. Each individual page was treated as a vertex in the graph. Hence the resulting community usually consists of a few hundred pages, some of which may from the same site. The member pages were ranked by the sum of their inbound and outbound links within the community.

Web communities defined on individual web pages are different from communities of weblogs. Weblog is more than a collection of web pages. Each weblog is attached to a particular author. We can consider the weblog community as a community of people with web presentations. The links between weblogs can be viewed as a communication evidence of the two people.

A number of special weblog search services are running on Internet. Technorati (www.technorati.com) was founded in 2002 and is now tracking 53.2 million blogs. It provides basic blog post search as well as a number of other statistics such as tag cloud and hottest topics to give users an overview of the ongoing topics. Technorati's search returns individual weblog posts. Google released a beta version blog search services (http://blogsearch.google.com/) in 2005. Majority of the query results from google BlogSearch points to individual entry, with a few highlighted, matched weblogs appearing on the top. None of the blog search engine provides community discovery services.

A blog community can be viewed a network of bloggers who write on similar topics and from time to time read, write and comment on each others' posts. The network of weblogs written by community members are connected through various types of hyperlinks. The blog community can be viewed as a special type of social network [12].

Marlow [15] identified four distinct subtypes of hyperlinks that can be viewed as weblog social ties. These are: blogrolls, permalinks, comments and trackbacks. Many bloggers list a few other blogs they read on the siderbar. The list is called a blogroll. Blogroll usually consists of blogs of similar interest. Bloggers often cite or reference other bloggers' writing on their own posts. This is called permalinks. Bloggers may respond to other's writing through other channels. They may leave a comment on the posting site. Sometimes, a commenter identifies self by a link to his or her own weblog in the comment body. This enables a crawler to build the connection between the two weblogs. In addition, some blog authoring tools provide "trackback" feature. This is a protocol that a responder can use to notify the author that he(she) has cited a particular article of the author in his(her) own blog. The notification is achieved by sending a ping message to the original blog entry, which will then update its trackback list to include the senders's URL. Trackback and permalink are two different formats of the same kind of information. Assuming that distinct hyperlinks convey slightly different social information, Marlow [15] examined the temporal patterns of the

blogrolls and permalinks. An interesting observation made by them is that blogroll and permalinks do tell different stories. There are weblogs that appear in the blogrolls of many weblogs but are not cited extensively. Marlow [15] suggested that permalink rank might be a more accurate way of measuring influence of a weblog than the blogroll.

Lin. et. al[13] adopted a similar social network perspective on weblog communities. They developed a measure called mutual awareness to quantify the weight of links between weblogs. One important feature of the mutual awareness measure is that it considers the time factor and assumes that the effect of actions on mutual awareness decays overtime. In [13] weblogs were organized in a graph with node representing the individual weblogs and edges indicating the ties between them. The strength of tie was calculated using the mutual awareness measure. It then utilized PageRank algorithm [3] to give each weblog a global ranking score and using the global ranking score and a set of predefined seeds, each of which representing a community, to perform clustering on the graph of weblogs. Lin et.al[13], conducted experiments on WWW2006 blogworkshop data set and NEC blog dataset and discovered a number of interesting communities with various structural patterns. The measures used by [13] and the rank scores calculated all require global information of the entire blogspace. Considering the dynamic feature of blogspace and possible scalability issue, we take a simpler measure, but operating on highly related subsets of blogspaces in constructing the graph and community extraction.

3. Blog Community Formal Definition

We consider each hyperlink as a communication instance between two bloggers. The strength of social tie is simply computed as the number of instances between bloggers. Blogroll link which may appear in different pages of a same weblog is considered as one instance. This would implicitly give more weight to permalinks. The scheme is consistent with the observation made in [15]. The temporal issue that may affect the measurement of social tie is discussed in section 6.

The general definition of blog community does not offer enough restrictions or details to construct any useful blog communities. The blogspace could grow to an unmanageable size and it may consist of many strongly connected sub-networks. We are more interested in those strongly connected sub-network rather than the loosely connected blogspace. We adopt Flake et. al.'s idea of web communities [6] and give a formal and operational definition of weblog community. In addition, we also define a few metrics to measure both the community and the individual weblog inside a community.

DEFINITION 1. A BLOGSPACE *is a network* $N = (V, E, w)$. $G = (V, E)$ *represents a connected directed graph, with each* $v \in V$ *represents a weblog and* $e_{s,t} \in E$ *represents a direct tie between two weblogs:* $s,t \in V$ *. For any given pair* (s,t) *There is at most one tie from s to t.* $w: E \to \Re$ *is a function assigning strength values to links.*

There will be a tie from weblog *s* to weblog *t* if and only if there is at least one of the hyperlinks discussed in section 2 from *s* to *t*. The simplest form of $w(s,t)$ is to take the sum of the number of hyperlinks from *s* to *t*.

DEFINITION 2. A BLOG COMMUNITY *is a vertex subset* $C \subset V$, *such that*

- */C/ <= N_{max} and /C/ >= N_{min} Where N_{max} and N_{min} are predefined maximum and minimum community sizes*
- *The corresponding Graph* $g \mid C = (C, L(C))$ *is connected, where L(C) represents the set of ties between vertices in C.*
- *And for all vertices* $v \in C$, *the weight of* v *inside* C *is larger than the weight of* v *to neighborhood. Here neighborhood refers to the set of vertices that have direct tie with* v *but are not included in* C *:*

$$\min p(v \mid C) > \max p(v \mid N(C))$$

This definition is a slightly modified version of the community definition given in [6]. It reflects the "community structure" observed in most social networks which states that there are groups of vertices that have high density of edges within them, with a low density of edges between groups[16]. Hence a simple and intuitive method of computing the weight of *V* inside and outside the cluster is to calculate the sum of direct tie strength values inside and outside the cluster.

We collect preliminary data on the evidence of the existence of such communities in blogspace. We download all relevant links as well as related information from a few weblogs and counted the frequency of different ties. A consistent pattern is that majority of links are weak ties with link weight around 1 or 2, only a small portion of links represent strong ties with weight over 10. It implies that the network of weblogs is not a purely random network and we can discover clusters based on the strength of ties. This conforms to Mark Granovetter's view of social world[8]. In his description, social world is structured

into highly connected clusters with a few external links connecting these clusters. In the following definitions, we use community to refer to blog community defined in DEFINITION 2.

DEFINITION 3 COMMUNITY CENTER *is defined as the weblog that has the highest betweenness score*.

The betweenness of a vertex in a graph is defined as the number of paths passing through this particular vertex. It measures the frequency with which a weblog falls between pairs of other weblogs on the shortest paths connecting them. It is one of the three classic measurements of point centrality in social network analysis. The intuition behind this measurement is that a vertex falls on the communication paths between other vertices will have a potential for control of their communication. The vertex that sits on the largest number of communication paths will play a central role in the whole graph. Freeman's measure of between score is adopted here [7]. Community center and betweeness scores are used to rank individual weblogs inside a community. It is a structural metric and serves similar purpose as the PageRank model in weblog community context.

DEFINITION 4. COMMUNITY CENTRALITY *is defined as* $C_B = \frac{\sum_{i=1}^{n}[C_B^{'}(v^*) - C_B^{'}(v_i)]}{n-1}$

Community centrality is a normalized metric to measure the extent a single weblog to be more central than all other weblogs in a community. Centrality is an important structural attribute of social networks. All previous social network research indicated that it is highly related with other important group properties and processes. Early empirical study showed that the speed and efficiency of a network in solving problems as well as the satisfaction of participants was related to the tendency of a single point to be outstandingly central[7]. It is the used to compare and rank communities of similar topics.

The metric given above is a direct application of the betweeness based graph centrality developed by Freeman[7]. It is computed as the average difference between the relative centrality of the most central point v^* and that of all other point. The most central point v^* is determined by its betweenness score. There are two other graph centrality measurements defined on degree centrality and closeness centrality as well. However, as observed by Freeman[7], betweenness is the most sensitive measurement among all proposed measurements. Small changes in topology and weight get largest variation in betweenness based centrality measure. The communities we need to rank here are of similar topics, and may interact occasionally with each other as they are all from a large blogspace. We should not expect huge structural difference among them. Therefore, betweenness based centrality is well-suited for the purpose of differentiating those kind of communities.

DEFINITION 5. COMMUNITY AUTHORITY *is defined as a weblog or weblogs that receive highest incoming tie strengths.*

DEFINITION 6. COMMUNITY HUB *is defined as a weblog or weblogs that have the highest outgoing tie strengths.*

Here we adapt the definition of authority and hub to make it refers to a weblog rather than a single page. Kleinberg also gave a simple iterative algorithm to discover the authorities and hubs in a network. In this algorithm, each vertex is supposed to a value of authority and a value of hub. The initial hub value is set to 1 for all vertices, the algorithm then takes turns to calculate corresponding authority value and hub value until an equilibrium is reached.

Both community hub and authority are important vertices in a community. The hub acts like a single point access to various sources, while the authority is like a very reliability source. Hub would be attractive to people who want to get a broad overview of the community discussion. On the other hand, authority would be a good place to easily identify the origin of the hottest discussion. Together with community center, we have introduced three important ranking methods in a community. Note that all of them are defined on the structural role. We are not dealing with content importance here. We will use real world experiment to check the validity and stability of the three measures.

4. Methodology

4.1 Blogspace acquisition

The starting point of weblog community identification is a connected blogspace with relatively similar topic. In practice, a search engine will keep a relative large copy of weblog pages and reconstruct the blogspace from the local copy. Since weblog rather than individual web page is the unit of analysis, we need specialized crawling and storage mechanism to support the weblog search. We develop a weblog crawler to construct desirable blogspaces.

The weblog crawler takes a weblog URL as seed and incrementally adds ties between weblogs in the collection. [15] identified four different sorts of hyperlinks pointing from source weblog to target

weblog as indicators of ties from source to target. The specialized crawler will collect those hyperlinks and store them as tie instances. To illustrate the idea, let us consider the simplest case with one source weblog, and suppose we are aiming to find all tie instances from the source to other weblogs. The basic algorithm has two steps:

1. Construct a large candidate set of hyperlinks from all web pages belonging to the source weblog.
2. Remove hyperlinks pointing back to the source weblog (including home page, entry pages, or possibly archive pages) and hyperlinks not pointing to weblogs. The remaining set consists of hyperlinks pointing to other weblogs. These hyperlinks may appear in the blogroll, entry body, comment or trackback section. Since we do not want to differentiate between different types, we consider each hyperlink in the remaining set an instance of some certain tie.

The basic algorithm will generate a star shaped graph with source in the center and all edges coming out from it. To construct the blogspace, we can repeat the basic algorithm for each distinct weblogs appear in the receiving end of the ties discovered so far. Table 1 illustrates the main crawling algorithm. In implementation, step 1 and step 2 are executed in pipeline style for performance and storage efficiency.

Table 1 The main weblog crawler algorithm

```
1    Url = seed url;
2    depth = 0;
3    method crawl (Url, depth)
4     if (depth < maxDepth)
5      for all hyper links link in Url
6       if link belongs to the same weblog
7            crawl (link, depth)
8       else if link points a weblog entry
9            find the home page of link as link.home
10           add a record Url.home and link.home
11           crawl(link.home, depth + 1)
12      end if
13     end for
14    end if
15   end method.
```

Depth is a variable to record the crawl depth. Seed is considered to have 0 depth. A weblog (including all pages belonging to it) that is on the receiving end of a tie from the seed will have a depth of 1 and those that are on the receiving end of ties from depth 1 weblogs will have depth of 2 and so on. It is used to control the size of the final blogspace. The *maxDepth* is set to 6 in most of the experimental runs to reflect the general "six degrees of separation" rule[1].

On step 6, a simple string prefix matching is performed to judge if a hyperlink belongs to source weblog. This is based on the observation that nearly all links belonging to a weblog have the same URL prefix determined by the URL of the weblog. For instance, all entries, archives or other pages within "The Old New Thing" weblog (http://blogs.msdn.com/oldnewthing/) will have a URL starts with "http://blogs.msdn.com/oldnewthing/".

On step 8, several rules are defined to determine if an extracted hyperlink pointing to a weblog entry. We first check if we can find any feed, be it RSS, ATOM or RDF feed from the page. If so, we consider the hyperlink a candidate Weblog link. Simple URL string match is used to rule out major news sites (News sites and weblogs are currently the two major users of feed).We next apply the algorithm developed by Ceglowski[4] to check if the link belongs to a weblog. Note that we have a slightly narrower definition of weblog than what is generally viewed as weblog. We add a restriction that all Weblogs need to provide feed in well-recognized format. It conforms to the trend that more and more newly added weblogs provide feed to let others syndicate their contents. The restriction also helps us to determine a blog entry's home page as explained in next paragraph. An exhaustive search for all possible feed formats is performed to look for feed on a webpage.

Given a feed URL, the homepage of a weblog can be retrieved from the feed content. The <link> value of the <channel> element in RSS usually gives the homepage address of a weblog. It enables us to crawl those weblogs with depth greater than 1 and use the homepage address to represent the whole weblog.

The blogroll list usually appears in homepage and every single entry page. To avoid counting them multiple times, the blogrolls were identified during the home page crawling and stored in a special place. They will be automatically removed from subsequent individual entry or archive page crawling.

The crawling result is a collection of relational records with two fields, the source weblog url and the target weblog url. Each record indicates a tie instance from source to target, with a value of 1. It is the relational representation of the resulting blogspace. One noteworthy feature of the resulting blogspace is that it does not contain self pointing edges. All edges are lines connecting two different vertices. Our crawl algorithm ensures that only relations with different source and target will be recorded. Self-pointing edges were ruled out because they do not confer any community related information and may affect the accuracy of the computation due to the effect of "inbreeding", that is, weblogs with large amount of internal links.

Starting from one weblog, we are able to get the complete list of weblogs this blogger reads, a complete list of other weblogs this blogger cites and all comments other bloggers made on his/her entries. They are considered as outgoing ties of a weblog. Links we can not get from a weblog includes comments this blogger leave in other weblogs and the collection of links that referencing or citing its posts (trackback features, if present, may give us small portion of this collection). These are considered as incoming links of that weblog. By setting appropriate crawl depth and crawling many interconnected weblogs, we can construct a partial list of those incoming links. From a blog reader's perspective, most of the incoming links are invisible and out of reach. We think it is justifiable to leave out some incoming links.

4.2 Blog community extraction

A single blogspace collected using the weblog crawler usually contains several thousands or more unique weblogs in the neighborhood of the seed. We expect to discover community structure from it. In constructing the directed graph, we first remove the multiple lines between two vertices and assign the line value as the number of lines. This is interpreted as the strength of tie between weblogs.

We adopt the island partitioning algorithm developed by Batagelj ([2]) to extract blog communities. Here island is defined as a connected small sub-network of size [*min, max*] with stronger internal line weight than line weight to vertices outside the sub-network. This corresponds to our Blog Community definition. Island partition is a hierarchical clustering method. The algorithm first orders all ties in decreasing order of their strength. It then, in the sorted order, merges ties to form sub-network, based on common vertex into sub-networks. All sub-networks form a hierarchical structure, with possibly a common root as the connected graph. Desirable sub-network were chosen from the hierarchy based on the size range. It is set to [5,50] in the experiment.

5. Experimental Results

We implemented the weblog crawler described in the previous section to extract weblog communities related to a few different seed urls. In this section, we reported results of two popular web services blogs selected from a list given in webservices.org. General themes and implications from all experiments will be discussed in section six.

5.1 Seed: Savas.parastatidist.name

The original blogspace constructed using Savas Parastadist's weblog as seed contains around 3800 unique weblogs and over 33000 records. After the initial pruning around 1500 vertices were left. We then apply the "island" partition algorithm to identify communities of size between 5 and 50 weblogs. We use Pajek[17] to perform the actual partition and the visualization. Figure 1 is a visual display of the resulting community using Fruchterman Reingold automatic layout generation algorithm.

Fifteen communities are identified from the blogspace. Except for four relatively big communities with around 30 members, the rest are small communities with around or less than 10 members. Table 2 gives the centrality index of the 15 communities. Most small communities have high centrality indices. This is obvious from the star shape topology of them. A star structure is defined to have the highest centrality score in SNA(Social Network Analysis). We next provide detailed inspections on two large communities: community 8, which contains the original seed; and community 10 which has the highest centrality index score among all large communities.

Community 8 contains the original seed (http://savas.parastatidis.name) and other 35 weblogs (jim.webber.name, which interacts frequently with the seed weblog, was down during the data collection period.). Figure 2 shows the members and the network structure. The topics of those members include .NET, xml and general web services. They are all related with web services. All members are weblogs except for node 2, which has a very similar structure with weblogs. Node 30 is itself a community on .NET and web services with many contributors. Most weblogs update frequently on their focused topics.

The calculated hub node 13 is actually a link collection weblog, almost every single post is a list of links related with .net, c#, and web services. The calculated hub node 28, however is not that prominent in the community. It has been referred 53 times by the hub thus has a higher authority score. However, it does not have any interaction with other members. This suggests a modified algorithm with refined line weight might be more appropriate here. The community center node 6 does have many incoming and outgoing links with other community member. This suggests that community center may be a more appropriate measure of the influence of a weblog in a community.

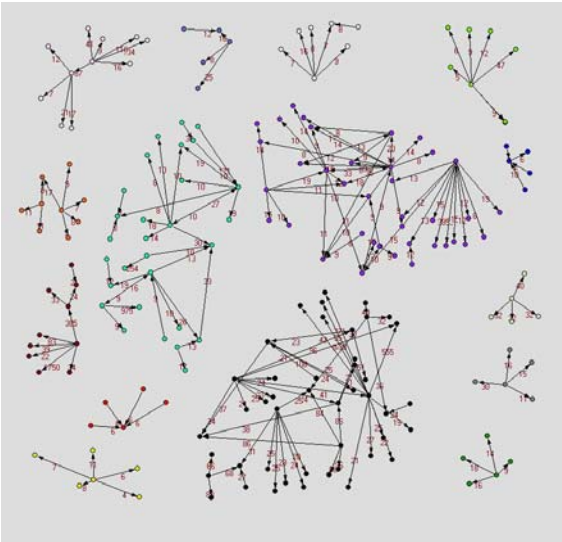
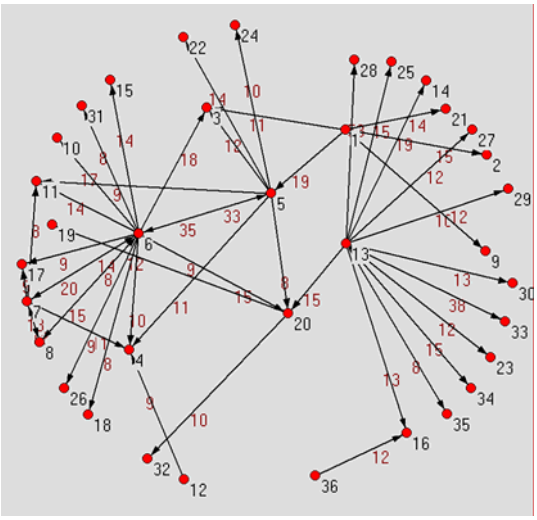


Figure 1 Web Services blog communities I

Table 2 Community centrality index

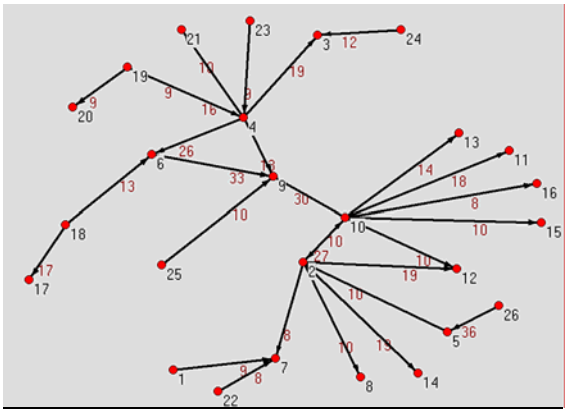
Community	Graph Centrality	Community Size
1	1	6
2	1	8
3	1	5
4	1	5
5	0.55	11
6	0.43	9
7	0.78	7
8	0.35	36
9	0.31	27
10	0.63	26
11	1	5
12	1	5
13	0.22	47
14	0.55	10
15	1	5



1	http://savas.parastatidis.name
2	http://research.microsoft.com/news/msrnews/
3	http://pluralsight.com/blogs/dbox/
4	http://pluralsight.com/blogs/craig/
5	http://pluralsight.com/blogs/tewald/
6	**http://pluralsight.com/blogs/aaron/** ***
7	http://pluralsight.com/blogs/keith/
8	http://pluralsight.com/blogs/fritz/
9	http://pluralsight.com/blogs/mgudgin/
10	http://msdn.microsoft.com/msdnmag/
11	http://unboxedsolutions.com/sean/
12	http://glazkov.com/blog/
13	**http://devauthority.com/blogs/csteen** *
14	http://weblogs.asp.net/ericjsmith/
15	http://blogs.msdn.com/smguest
16	http://samgentile.com/blog/
17	http://weblogs.asp.net/rhurlbut/
18	http://jcooney.net/
19	http://blogs.msdn.com/yassers
20	http://weblogs.asp.net/cweyer/
21	http://www.innoq.com/blog/st/
22	http://www.jonfancey.com/
23	http://codebetter.com/blogs/jeffrey.palermo
24	http://blog.whatfettle.com/
25	http://blogs.msdn.com/brada
26	http://blogs.msdn.com/mpowell
27	http://weblogs.asp.net/jgaylord/
28	**http://blogs.msdn.com/robcaron** **
29	http://weblogs.asp.net/despos/
30	http://www.theserverside.net
31	http://blogs.msdn.com/mfussell
32	http://weblogs.asp.net/mnissen/
33	http://blogs.msdn.com/trobbins
34	http://blogs.msdn.com/tomholl
35	http://weblogs.asp.net/wallym/
36	http://dotnetjunkies.com/WebLog/barblog
*	The calculated hub
**	The calculated Authority
***	The calculated community center

Figure 2 Community 8 structure and members

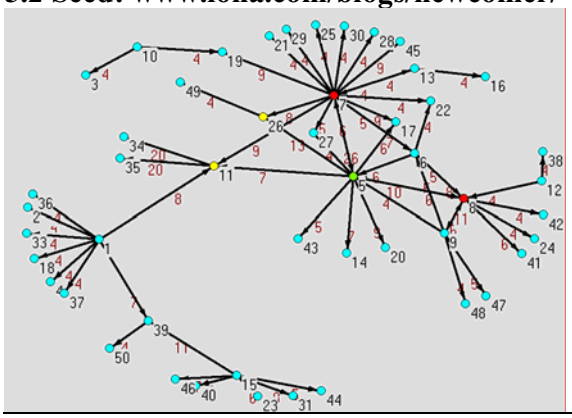
Figure 3 illustrates community 10 and its most important members. In this case the hub is node 10 which has many outgoing links; authority is node 9 which has quite a few incoming links from various vertices, and center is node 2 which sits in many communication paths between other weblogs. This community is more focused on .NET technologies rather than general web services.



Hub	Node 10: http://codebetter.com/blogs/raymond.lewallen
Authority	Node 9: http://codebetter.com/blogs/darrell.norton
Community center	Node 2: http://codebetter.com/blogs/sahil.malik

Figure 3 Community 10 structure

5.2 Seed: www.iona.com/blogs/newcomer/



Hub	Node 7: http://pluralsight.com/blogs/tewald/
	Node 8: http://pluralsight.com/blogs/keith/
	Node 5: http://pluralsight.com/blogs/aaron/
Authority	Node 26: http://weblogs.asp.net/cweyer/
	Node 11: http://pluralsight.com/blogs/dbox/
	Node 5: http://pluralsight.com/blogs/aaron/
Community center	Node 7: http://pluralsight.com/blogs/tewald/

Figure 4 Community 2 structure

The blogspace constructed using www.iona.com/blogs/newcomer/ as seed contains around 700 unique weblogs and over 3000 links. Three different communities were discovered from this blogspace. Figure 4 illustrates the community which contains the seed weblog and illustrate the important vertices in it. The main theme of this community is also dot net technologies. We calculate 3 hubs and authorities instead of one for this community and we also use the betweeness score to calculate the center of the community. They are illustrated using different colors in the graph.

6. Discussion

6.1 Quality of the Results

The experiments demonstrate that our algorithm can extract communities with focused themes related with a particular topic. Starting from a general web services weblog, we have discovered communities of general web services, of .NET and other web services related products. These can hardly be achieved through keyword search with limited semantic enhancement. In addition, majority of the community members we discover are not able to be located through current blog search engines. We used the keyword "web services",".NET","ASP.NET" in Google's BlogSearch, bloglines.com, blogdex.net and daypop.com and checked the top 20 results returned. Except for blogdex.net, which contains some web services weblogs. none of the rest has those members returned.

6.2 Seeding Effect

Our algorithm uses a small portion of the whole weblog space to discover communities discussing similar topics. The subset is obtained through specialized crawling starting from a seed URL. The selection of seed may affect the efficiency and quality of the communities returned. In one of our experiment we used Amazon Web Services (http://aws.typepad.com/aws/), one of the top weblog results for term "web services" from google's blogsearch service. The initial blogspace contains around 8000 unique weblogs and over 50000 links. Only one valid community is discovered from the partitioning algorithm. AWS itself is not included in the community and the community is not particularly focused on web services discussion. This result has two implications. First, AWS does not have an active community around it. Second, AWS is not within six clicks distance to any active web services blog community.

The proposed algorithm is useful for discovering weblog communities around a particular weblog or a few clicks away from a particular weblog. Using keywords as input parameter for community discovery will need a preprocessing to identify the suitable seed.

Another minor seeding effect observed in our experiment is that blog communities discovered from different blogspaces may have overlaps. Note that a few *pluralsight* members (for instance, http://pluralsight.com/blogs/aaron/) appeared in Figure 2 and Figure 3 communities. Their rankings with respect to structural scores (centrality score, hub score or authority score) may be different in different communities. This happens because each blogspace represents a partial view of the whole weblog space. In this particular experiment, it represents the six degree neighborhood of a particular weblog. Increasing the degree will minimize this problem.

6.3 Clustering Feature

A recurring theme in the community construction is that many communities are formed by members from the same hosting services. For instance, in the community 10 illustrated in figure 3, one half of the members are from the *codebetter.com* hosting service. The community in figure 4 has large number of members from both *codebetter.com* and *pluralsigh.com*. Apart from these two, we also discover communities consisting largely of members from *blogs.msdn.com, radio.weblogs.com* and other hosting services in some of our experiments.

6.4 Metrics Evaluation

We define one community wide metric and three individual level metrics to measure the prominent features of a weblog community. The graph centrality index measures the extent of a particular weblog to be the sole center of a community. It is currently biased towards small communities with star structure. We may need to refine the weight on community size to ameliorate the score.

For each community, three important vertices are identified: the hub, the authority and the center. The center is determined through a vertex's betweenness score and can always accurately represent the vertex that sits in most communication paths. Kleinberg's HITS algorithm is used to determine the hub and authority vertex. In this algorithm, the hub and authority value have a mutual dependency on each other and may subject to calculation bias as exemplified by the web services community in figure 2. So far, community center is the most stable measure in member ranking.

6.5 Historic Entry Data Issue

The blogspace we use as base set contains entire historic data from a weblog. If a weblog has been running for several years, social ties created during all those years will be considered in community discovery. This may raise argument that too much historic data may affect the accuracy of clustering and partitioning of the blogspace as well as the identification of the important nodes in a community. For one thing, readers probably are not interested in early writings; if they start reading a weblog on 2005, it is hardly the case that they will want to read a 1999 entry. For another thing, recent social interaction patterns could be different from the historic ones and in that case would be buried among huge amount of historic data.

To develop a better understanding of the problem, we carry a micro-scale case study on a few weblogs. We expect to get some preliminary idea of the average life span of a blog entry. We chose two weblogs with heavy comments and downloaded all comments along with the time they were made as well as the time the corresponding entry was created. We use the latest time a comment was made by somebody other than the author and the blog posting time to represent the life span of a blog entry. Two weblogs http://blogs.msdn.com/ericlippert/ and http://www.micropersuasion.com/ are chosen as the sample. Both have been running for a few years and have considerable number of comments. Around 5000 comments are downloaded from the 605 entries of ericlippert weblog and over 8000 comments are downloaded from the 1384 entries of micropersuassion weblog. The life span of entries varies significantly with most entries having a one day life span and a few having life spans as high as nearly two years. For instance, one entry in Eric Lipert's blog (http://blogs.msdn.com/ericlippert/archive/2003/10/06/53150.aspx) originally posted in November, 2003 attracted some recent discussion in September, 2005, around two years later. There are quite few posts in Eric's weblog with more than 600 days of life span. The average entry life span is around 6 days in Micropersuassion with a standard deviation of 36.7 days. The average entry life span is around 75 days in Eric Lipert's weblog with a standard deviation of 168.9 days. Figure 5 gives the frequency distribution of the entry life span in these two weblogs. In both cases, the distribution follows Zipf-like curve. Although we can not generalize the findings from these two case studies, they surely offer evidence, which suggests that

historical entries are not completely ignored by blog readers.

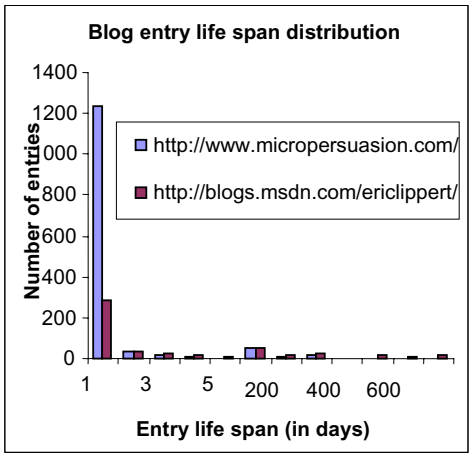


Figure 5 Blog entry life span

On the other hand, majority of the entries have very short life span, suggesting that focusing on recent data may get better results to most users' interest. In that case, the data acquisition method can be adapted to build blogspace with recent data. We can apply the algorithm on feed data rather than on entire weblogs and use a sliding time window to update the blogspace with new feeds. However, we believe that there will not be a standard window size that can fit all sorts of welbogs. Different weblogs usually have very different average entry life span.

7. Conclusions

In this paper, we gave a formal definition of weblog community based on social network theories and observations. We also defined a few metrics to describe the feature of a community as well as the important nodes in it.

There are many applications for the methods we proposed. The specialized web crawler suggests new ways of handling and storing weblog related data with an emphasis on the social tie information. It can be used directly to prepare data for community discovery. Alternatively, it can be used to search local copy of web pages to build the blogspace. The metrics developed can be used to rank communities as well as weblogs inside communities.

8. References

[1] Barabasi A.L., Linked: the new science of networks.

[2] Batagelj V. Analysis of large networks – Islands Presented at Dagstuhl seminar 03361: *Algorithmic Aspects of Large and Complex Networks* Dagstuhl, 2003

[3] Brin S and Page L, The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW1998*, 1998, Brisbane, Australia

[4] Ceglowski.M. Www::identify-identify blogging tools based on url and content. http://search.cpan.org/~mceglows/WWW-Blog-Identify-0.06/Identify.pm, 2003

[5] Fitzgerald M. "Case Study: Going Up Against Google." *Inc. Magazine*, Feburary 2006, [online document] [cited 01.09.2006] Available at http://www.inc.com/magazine/20060201/handson-casestudy.

[6] Flake G.W., Lawrence S, and Giles C.L. Efficient identification of web communities. In *Proc. 6th ACM SIGKDD Intel. Conf. On Knowledge Discovery and Data Mining*, 2000, page 150-160

[7] Freeman L. C. Centrality in Social Networks Conceptual Clarification, *Social Networks* 1(1978/79), 215-239

[8] Granovetter M. The Strength of weak ties:a network theory revisited. *Sociological Theory*, Vol. 1 (1983), 201-233

[9] Kleinberg J. Authoritative sources in a hyperlinked environment. *Journal of the ACM* Vol.46 No.5 1999: 604-632

[10] Kumar R. Novak J., Raghavan P. and Tomkins A. On the bursty evolution of blogspace. *WWW2003*, May 2003, Budapest, Hungary

[11] Kumar R., Raghavan P., and Tomkins A. Trawling the Web for emerging cyber-communities. In *Proc. 8th WWW Conference, 1999*

[12] Lento T., H. T. Welser, L. Gu and M. Smith, "The Ties that Blog: Examining the Relationship Between Social Ties and Continued participation in the Wallop Weblogging System," In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystems:Aggregation, Analysis and Dynamics*, *WWW2006*, Edinburgh, May 23, 2006

[13] Lin Y. , H. Sundaram, Y. Chi, J. Tatemura and B. Tseng, "Discovery of Blog communities based on Mutual Awareness", In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystems:Aggregation, Analysis and Dynamics*, *WWW2006*, Edinburgh, May 23, 2006

[14] Meislin Rich, Blog 101. http://www.nytimes.com/ref/technology/blogs_101.html

[15] Marlow, C. 2004. Audience, Structure and Authority in the Weblog Community. Presented at the International Communication Association Conference.

[16] Newman M.E.J. The Structure and function of complex networks, *Condensed Matter* V0l. 1, 0303516, 2003

[17] Pajek, http://vlado.fmf.unilj.si/pub/networks/pajek/

[18] Toyoda M., Kitsuregawa M. Extracting Evolution of Web Communities from a Series of Web Archives. *Hypertext*, 2003, Nottingham, UK.