Discovery of Blog Communities based on Mutual Awareness

Yu-Ru Lin Hari Sundaram Yun Chi Jun Tatemura Belle Tseng
Arts Media and Engineering Program
Arizona State University
NEC Laboratories America
Cupertino, CA 95014
e-mail: {yu-ru.lin, hari.sundaram}@asu.edu, {ychi, tatemura, belle}@sv.nec-labs.com

Abstract

Blogs have many fast growing communities on the Internet. Discovering such communities in the blogosphere is important for sustaining and encouraging new blogger participation. We focus on extracting communities based on two key insights – (a) communities form due to individual blogger actions that are mutually observable; (b) semantics of the hyperlink structure are different from traditional web analysis problems. Our approach involves developing computational models for mutual awareness that incorporates the specific action type, frequency and time of occurrence. We use the mutual awareness feature with a ranking-based community extraction algorithm to discover communities. To validate our approach, four performance measures are used on the WWW2006 Blog Workshop dataset and the NEC focused blog dataset with excellent quantitative results. The extracted communities also demonstrate to be semantically cohesive with respect to their topics of interest.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: *Information Search and Retrieval*; H.3.5 [**Information Systems**]: *Online Information Services*; H.5.4 [**Information Systems**]: *Hypertext/Hypermedia*

Keywords

Blogs, community, community extraction, mutual awareness, ranking, clustering, social network analysis.

1. INTRODUCTION

Recently, blogs (or weblogs) have become prominent social media on the Internet that enable users to quickly and easily publish content for their community. In the blogosphere, various communities are formed since the blog encourages users to communicate with each other through various facilities such as comments and trackback.

The blog community structure emerges through individual bloggers' behavior – how bloggers read and communicate ideas with other bloggers. The blogger becomes both the *producer* of content (author of her own blog) as well as the *consumer* who reads other blogs and web pages for potentially interesting events that she may later blog about.

To form a community, it is critical that individual bloggers become aware of each other's presence through interaction. We refer to this bi-directional property as "mutual awareness" of bloggers. A blog community that is formed based on mutual awareness is very different from communities defined in the traditional Web analysis literature [3,7,10] since the semantics of the hyperlink structures are different. In the Web structure analysis, it is commonly assumed that if a page links to another page, the two are related to each other. A community of web pages is formed based on this "relevance." On the other hand, a blog community is based on mutual awareness amongst bloggers, which is only present as a result of bi-directional communication. Figure 1 shows two graphs of bloggers with unidirectional communication. Since no blogger is aware of others' actions, these graphs should not be considered as communities.

Mutual awareness is established through the various actions of bloggers. While some actions can directly lead to both bloggers becoming aware of each other (e.g., comments, trackback), other actions do not (e.g., entry-to-entry links).

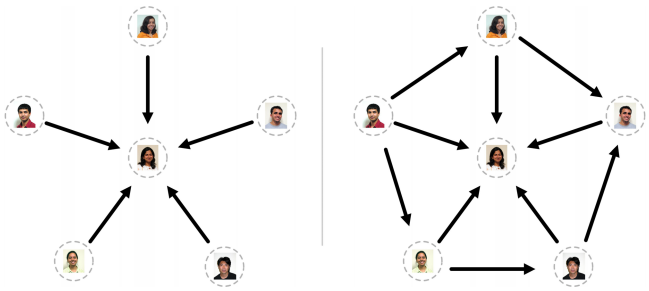


Figure 1: Both figures show only *uni-directional* communication between bloggers (e.g. creating entry-to-entry links). Bloggers *are not aware* of each other, unless the center blogger sends a trackback. Thus graphs without bi-directional communication cannot form a desirable community as there is no mutual awareness amongst the members of the group.

Based on the above insights, we propose a new approach to community extraction that consists of two steps – (a) analysis of mutual awareness from bloggers' actions and (b) ranking-based community extraction from mutual awareness. Mutual awareness between two bloggers is affected by the type of action, the number of actions for each type, and when the actions occurred. For each action type, we compute an *action matrix*, where the entries are proportional to the number of action occurrences and exponentially weighted by their time of occurrence. We compute the *mutual awareness matrix* as a weighted linear combination of action matrices. Note that the coefficient is a symmetric, bi-directional property for a pair of bloggers. We use an iterative, ranking-based clustering scheme on the mutual awareness matrix to determine the communities. PageRank [3] is used to determine the seeds at each step, followed by diffusion [15] of association to determine the community members.

We have conducted extensive experiments on two different datasets, with excellent results. The first is the WWW2006 Blog Workshop dataset, and the second is an NEC Blog dataset, that has a strong technology focus spanning over six months. We have used the following four validation metrics to analyze our results – (a) coverage, (b) conductance, (c) interest coefficient, and (d) sustainability. The first two are standard graph theoretic metrics and serve as baseline comparison, while the last two are

validation metrics introduced by us to better represent the dynamic and temporal properties of the blogs. Our community extraction results indicate that the mutual awareness matrix outperforms the traditional adjacency matrix on all performance measures, and extending over both datasets. The detected communities also show a strong topical cohesiveness. Even in the absence of ground truth, these results are strongly suggestive of the value of our mutual awareness approach.

The rest of paper is organized as follows. In the next section we discuss related work on community formation theory, and computational approaches to community extraction. In section 3, the unique properties of blogs suggest the need for a new analytical framework. In section 4, we highlight our approach towards community formation. Section 5 presents the computational models for mutual awareness and derives an algorithm for community extraction. In section 6, four metrics are identified for validation, and in section 7, detailed experimental results are illustrated for the two datasets. Finally, we present our summary and conclusion in section 8.

2. RELATED WORK

In this section, we discuss related research work for the definition and extraction of communities. In general, the identification of communities involves both sociological and technological issues. From sociology, community formation is a social process where different notions of community are used depending on the social context. From information technology, different engineering solutions exist for community extraction based on corresponding specifications of community. Therefore, we review related research work both on the notion of a community and on community extraction techniques.

2.1 The Notion of a Community

The notion of *virtual community*, or online community, has been discussed extensively in prior research. Rheingold [12] defined virtual communities to be *"social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationship in cyberspace"*. Jones [9] further clarified the notion of virtual community based on the definition of *virtual settlement*—the place, or cyberplace, where a virtual community forms—and the characteristics of virtual settlements. Jones considered four characteristics of virtual settlement as the necessary conditions for the formation of a virtual community: (1) interactivity, (2) communicators, (3) virtual common-public-place where the computer-mediated communication takes place, and (4) sustained membership. The interactive nature of virtual community distinguishes a (virtual) community from a *group*—a virtual community is not a chance meeting of casual individuals but should involve long term, meaningful conversations among humans, and this condition suggests that there should be a minimal level of sustained membership. The idea that interactivity forms a social reality has also been discussed by Dourish [5]. According to Dourish, interaction involves presence (some way of making the actors *present* in the locale) and awareness (some way of being *aware* of the other's presence). In what Dourish called an *action community*, members share the common sense understandings through the reciprocal actions.

Blanchard [1] in part extended the work of Jones [9] to analyze the sense of community among blogs. Blanchard argued that sense of community is an essential characteristic that distinguishes virtual community from mere virtual group. Without specifying what the sense of community is, Blanchard conducted a survey of the blog readers to examine the existence of virtual community in blogspace. The analysis focused on the interactivity among a prestigious blogger and his or her audiences and studied possibilities of interaction among the audiences.

Although prior studies have highlighted the importance of action and interaction in human communities, there has been little work using theoretical frameworks of action to detect blog communities.

2.2 Community Extraction

The extraction of communities is mainly studied as a graph problem. Individuals and the relationships between individuals are represented as nodes and edges on a graph, respectively. The problem of community extraction is transformed into grouping cohesive nodes in terms of their relationship defined as a function on the edges.

One research area related to our work is graph partitioning. In the graph partitioning problem, the goals are to partition the set of nodes on a graph into disjoint subsets so that (1) the association among nodes within the same subset is high and (2) the disassociation among nodes across different subsets is low. Shi et al. [13] showed that the above two goals can be satisfied simultaneously by using appropriate definitions of association and disassociation (they used *normalized association* for the former and *normalized cut* for the latter). Shi et al. further showed that under their definitions, although the graph partitioning problem is NP-complete, the problem can be converted to a problem of optimizing a specific Rayleigh quotient and then relaxed to solve a generalized eigenvalue system, which can be solved efficiently. Following similar ideas, Dhillon et al. [4] proved that the objective of spectral clustering methods used in graph partitioning is mathematically equivalent to that of a general weighted kernel k-means problem. Because of this equivalence, one can choose desirable existing solutions for either problem to solve the other problem.

We argue that the purpose of the graph partitioning problem is different from the problem of community extraction in the blogosphere. In the graph partitioning problem, every node is equally important and has to been assigned to one of the partitions. In contrast, we consider the majority of the bloggers in the blogosphere to be unimportant and we instead focus on a small fraction of bloggers that form dense groups. Our view of community in the blogosphere is in line with other research work in the area of Web and blog analysis. For instance, Flake et al. [6] defined a Web community as "a set of sites that have more links to members of the community than to non-members". That is, a community is a subgraph of the Web graph that is dense in a certain sense. Flake et al. also proposed algorithms for identifying such communities by using a maximum flow/minimum cut framework. Kumar et al. [11] and Ino et al. [8] proposed different algorithms based on ideas similar to that of Flake et al. However, our proposed community extraction algorithm is different from the above algorithms in two ways. First, our algorithm discovers dense subgraph by using global linkage information instead of local topological structures. Second, instead of one simple blog graph, our algorithm uses a different graph that encodes intrinsic relationships among bloggers where the relationship is derived from different types of links in the blogosphere that are observed at different time.

Our community extraction algorithm was partly inspired by two algorithms. In the first work, Zhou et al. [15] proposed to modify

the random walk model in PageRank [3] to solve the semi-supervised learning problem, which seeks to classify all points into labeling groups when only a few labeled points are available. In the second work, Tseng et al. [14] proposed to visualize blog communities by a tomographic clustering graph where the height of a blog in the graph is a score extended from the PageRank score. In both work, membership is propagated from a given node to other nodes whose memberships have not been determined. We use a similar idea in our community extraction algorithm.

To summarize, existing research focuses on either using global link structure to search for an exhaustive graph partition, or using only local connection information for a dense subgraph extraction. In addition, most existing community extraction methods analyze the Web or the blogosphere as a static graph. Furthermore, the above related research work used only directly observable communications among individuals and ignored the inherent hidden social relationship and mutual awareness among individuals that caused the observed communications.

3. ANALYTICAL FRAMEWORK

Community extraction requires a new analytical framework that incorporates the unique properties of blogs. This framework must be grounded in how bloggers' act in the blogosphere, and the consequences of their actions. Figure 2 illustrates how the interactions between two bloggers convey a sense of mutual awareness. For this new medium, let us examine four unique properties of blogs and how they differentiate from the Web.

1. **Temporal dynamics:** A blog differs in a fundamental way from regular web pages due to its temporal nature. Blogs represent easily editable (i.e. online, rather than using a specialized program) content where both the author and people reading the blog can add content, through entries and comments respectively. This editable characteristic has led blogs to become highly dynamic media where people create new content on a regular basis.

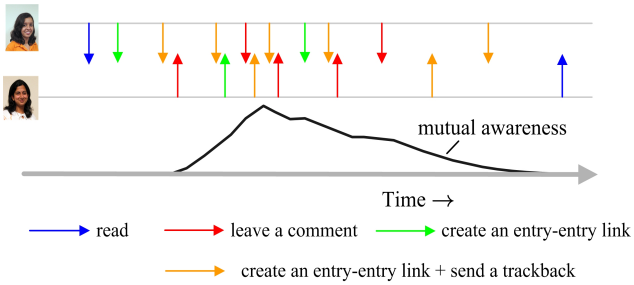


Figure 2: Mutual awareness between two bloggers is affected by the type of actions, the number of such actions, and when such actions occur. The arrow direction indicates the source and the destination blogger on whom the action is performed. A sample mutual awareness curve is plotted to show the action impacts.

2. **Event Locality**: A typical blog entry is time sensitive – i.e. it is well correlated to either global events, or to events that are observed by the author. This implies that the relevance of the entry to the author as well as the entry context are highly correlated with the time that it was created. The author will not explain the context directly, but will assume its existence when creating the entry. An entry that is read say five years later may lack the context for it to be understood.
3. **Link Semantics**: A hyperlink can have different semantics depending on the context of its use. For example, URL's on blogrolls are very significant to the blogger, as they represent a public display of interest. Bloggers are very careful when they add / delete URL's that are part of the blogroll. Trackback links have a different effect than regular entry to entry links as they reveal as they make both bloggers aware of each other. An entry to entry link is a hyperlink embedded in the content of the blogger and pointing to the other blogger's entry. It is unidirectional in sense that the blogger whose entry / blog has been linked to, will not become aware of the link.
4. **Community Centric**: In the blogosphere, the dynamic content creation process is sustained by human action and interest. The blog lets our peers easily provide feedback on our content either directly (e.g. comments, trackbacks) or indirectly (e.g., entry-entry links, blogrolls) leading to the formation of communities - groups of people who are interested in each others' content.

We note that the community extraction problem is distinct from traditional ranking problems on the web [3,7,10]. A key difference lies in the semantics of the hyperlinked structure. These differences in the semantics are easily understood using a simple graph structure as shown in Figure 3. Let us assume that the blogger at the center plays the role of a hub – she creates entry-to-entry links to other bloggers, whose entries she likes. This results in a familiar graph structure – a star network. Let us further assume that the bloggers thus connected to the blogger at the center, are *not* connected to each other via any hyperlinks.

We claim that this star network is not a community – it does not represent a group interested in each other's content. An entry-to-entry link is not observable by the bloggers whose entries have been thus linked, unless the blogger at center of the figure sends each one of them a trackback. Thus this network cannot be a community – there is no mutual awareness. A similar argument can be made for authoritative bloggers whose entries have been linked to by many other bloggers. This too is a star network, but the bloggers at the periphery are unaware of each other, and the blogger at the center is unaware that many others have linked to her. In both examples, the communication was unidirectional. *In order for communities to form, bi-directional communication leading to increase in mutual awareness is critical.*

Hubs and authorities play an important role in web search algorithms [3,10] that seek to determine the most important information source for a specific query. The intuition behind algorithms such as HITS [10] or PageRank [3] lies in interpreting the *hyperlink structure as the navigation path for a web surfer*. Hence the emphasis is on the analysis of the in-degree and the out-degree edge distributions for a node (web page or website) to determine the best information source.

In the blogosphere, communities emerge through the sustained action of *individual bloggers, not through the navigation of casual web surfers.* In the blogosphere, the blogger becomes both the creator of content (her own blog) as well as a focused surfer who looks at other blogs / website for potentially interesting events that she can later blog. Communities are important to the blogger who wishes to participate in them through a dialogue with

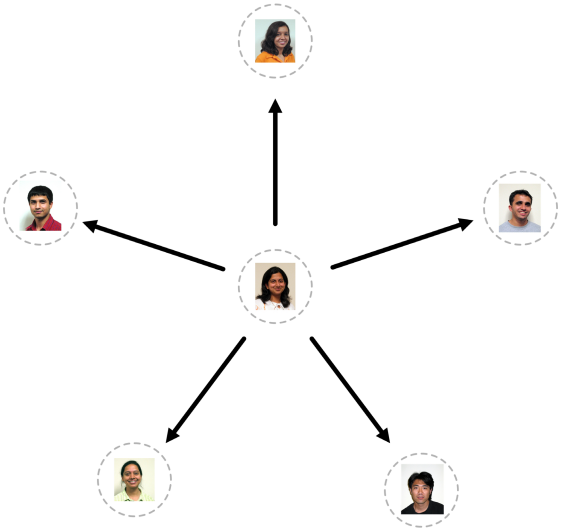


Figure 3: The center blogger plays the role of a hub, by creating entry-to-entry links. This results in a familiar graph – the star network. Bloggers on the periphery are not aware of each other, unless the center blogger sends trackbacks. Thus this graph is not a valid community since there is no mutual awareness amongst the members.

fellow bloggers. However, the casual web surfer may be only looking for specific content and not developing a relationship with other people.

We now propose a mutual awareness framework for community extraction that incorporates the rich semantic, temporal and community aspects of the blogosphere.

4. COMMUNITY FORMATION

The blog community structure is developed through individual blogger's behavioral pattern—how bloggers read and communicate ideas with other bloggers. The bloggers can act in the blogosphere, in the following ways (a) surf / read (b) create entries (containing entry-to-entry links, entry to blog / web, or no link), (c) comment and (d) change blogroll. The last three actions are observable through the data collected from the blogosphere, but the first action is hidden.

Mutual awareness of individual blogger action is critical to community formation. By mutual awareness of action we mean that individual blogger actions must lead to bloggers becoming aware of each others presence – *this is a bi-directional property*. These ideas are influenced by Locale theory [5] that discusses how social organization of activity is supported in different spaces. While the domains of activity must provide means for the community members to act, the space must also accord members' *presence* and facilitate *mutual awareness*.

It is interesting to note that some blogger actions are not observable by other bloggers. For example, let us consider two hypothetical bloggers – Mary and John. Let us assume that Mary creates an entry with a hyperlink that points to John's blog. In this case John would be unaware of Mary's entry. On the other hand, if Mary leaves a comment on John's entry, then John is immediately aware of her presence. Clearly, mutual awareness is a critical aspect of community formation.

We now examine different blogger actions in terms of mutual awareness.

- *Create Entries*: If Mary creates a new entry that has (1) no hyperlink, (2) a link to the web (that John may have also linked to), or (3) a link to John's blog, then the action cannot lead to mutual awareness. For an entry that contains a hyperlink to John's entry, the action can lead to mutual awareness, provided that Mary sends John a trackback link. Note that if the blog data shows that Mary and John have created multiple entry to entry links to each others blog entries over a period of time, it is likely that they have become aware of each other. This is because both would have read the other's entry that contains a link to their own entry.
- *Comment*: This action leads to increase in mutual awareness. For example, Mary reveals her presence to John, by leaving a comment on his blog. Note that she cannot be sure that John has actually read her comment, unless John acts in an observable manner to indicate so (e.g. commenting on Mary's blog; responding to Mary's comment etc.)
- *Change Blogroll*: Addition of a new link to John's blog by Mary, to her blogroll is a significant event, as the blogroll serves as a public declaration of affinity. However, again John cannot become aware of her presence in the blogosphere, unless he visits her blog by chance.

The impact of a specific blogger action depends on two factors – (a) if the action can lead to mutual awareness, and (b) the importance of the action for the blogger who performs the action. If Mary *mostly* leaves comments on other bloggers, and the importance of a comment for Mary is low – while many bloggers are aware of Mary, she may not feel that she in engaged in dialogue with them.

In this section we have presented the idea that communities form due to individual blogger actions that lead to mutual awareness. A community needs to be *sustained over time* though individual action (i.e. mutually observable actions of individual bloggers over a period of time). This is the important distinction with a group of people who may have a chance encounter with each other.

5. EXTRACTING BLOG COMMUNITIES

In this section, we present an approach for extracting blog communities. The key idea of our approach is to (1) compute the *mutual awareness* as edge weight in a graph-based representation of blogspace, and then (2) use a ranking-based clustering method to extract blog communities.

5.1 Computing Mutual Awareness

Mutual awareness between two bloggers is affected by the type of action, the number of actions for each type, and when the action occurred (ref. Figure 2). It depends on *sustained* actions — it increases if there are follow-up actions that lead to mutual awareness and decreases if actions are not sustained over time.

5.1.1 Mutual awareness matrix

We now show how to compute a mutual awareness matrix for the members of the blogosphere. The set of bloggers in the blogspace are represented as a weighted directed graph $G = (V, E)$, where the nodes of the graph are the bloggers, and an edge between any pair of nodes i and j is a connection from i to j. The weight on

each edge w_{ij} is a function of mutual awareness. The graph can be represented as a matrix M, the *mutual awareness matrix*. We compute the matrix by both accounting for the specific actions, and the time when the action occurred.

Let A represent the set of actions possible by any individual blogger (ref. Figure 4), and a_k represent the k^{th} action type (e.g. a comment, or the creation of an entry-to-entry link, is a specific example of an action). For each action type k, and at time t, we can compute a temporal action matrix $X_{k,t}$. Each entry $x_{ij,k,t}$ of this matrix represents the number of times the k^{th} action a_k was performed by blogger i on blogger j (e.g. blogger i leaves a comment on blogger j's entry.). Note that the self action value (i.e. the diagonal elements) is zero and the matrix in general is not symmetric.

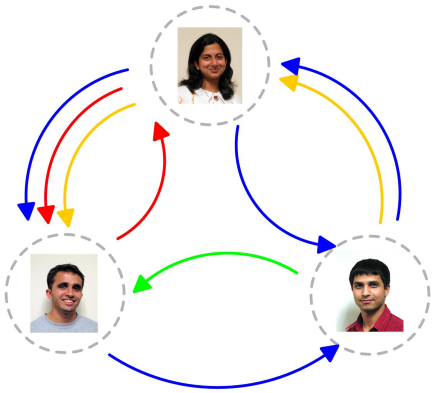


Figure 4: A simple three blogger interaction. The color of the edge represents a specific action type, and the direction of the arrow indicates the recipient of the action. In this graph all communications are bi-directional.

The effect of actions on mutual awareness decays over time. Mutual awareness due to earlier actions will gradually diminish. We model the effect of actions as a decaying exponential function. It is then straightforward to incorporate both effects. The aggregated action matrix X_k during a period of time between the initial time t_0 and current time T is computed as follows:

$$X_k = \sum_{t=t_0}^{T} X_{k,t} e^{-\lambda_k (T-t)}, \qquad <1>$$

where λ_k is the decaying factor for the action type k and is in general different per action. The effect is for different types of actions to decay at different rate.

Each type of actions can in general have different effects on mutual awareness. We now examine two specific types of actions, (a) create an entry-to-entry link and (b) send a trackback.

Create an entry-to-entry link: For action of creating an entry-to-entry link, the mutual awareness derived from this action is computed as follows:

$$M_k = \min(X_k, X_k^T), \qquad <2>$$

where M_k is the mutual awareness matrix due to the k^{th} action alone. We set $m_{ij} = 0$ if $m_{ij} < \lambda_m$, where λ_m represents a minimal level of mutual awareness.

Equation <2> implies that for an entry-to-entry link to give rise to mutual awareness, it must be reciprocal (i.e. i and j must both create entry-to-entry links to each other), and be above a certain threshold. This is because the action of creating an entry-to-entry link is not reciprocal—given a link from blog i to blog j, there is no guarantee that blogger j is aware of i. The reciprocity condition in Eq.<2> gives the possibility of both bloggers being aware of each other to be high.

Send a trackback: The action of sending a trackback will immediately lead to both bloggers becoming aware of each other. A trackback is a notice sent by a blogger i to another blog (or blogger) j to notify that i has read and cited j's entry. We compute the mutual awareness of this action by:

$$M_k = rX_k + (1-r)X_k^T, \qquad <3>$$

where r denotes how likely the trackback receiver is to be aware of the trackback sender by the action of sending a trackback.

5.1.2 Fusion of different Actions

If we assume that the actions a_k in the action set A are independent of each other, then we can consider linear combinations to fuse the information from both sources. We can specifically combine the mutual awareness matrices for both actions as follows:

$$M = \alpha_1 M_1 + \alpha_2 M_2, \qquad <4>$$

where the indices of 1 and 2 are used to denote the actions of entry-to-entry link creation and sending trackback respectively, and where α_1 and α_2 represent the weights for the mutual awareness matrix from each action.

In general, the composite mutual awareness matrix M combines M_k by awareness factor f_k for action k as follows:

$$M = \sum_k \alpha_k M_k \qquad <5>$$

Mutual awareness is a *bi-directional* relationship indicating how well a pair of bloggers is aware of each other. This semantics results in a symmetric mutual awareness matrix. We then use this matrix to identify the blog communities.

5.2 Ranking-Based Clustering Method

We use a ranking-based method to extract blog communities from the graph representation of blogosphere. In contrast to general graph partitioning problems that divide all the nodes into groups, we are only interested in groups of bloggers who are actively communicating with each other. Since the observable links such as entry-to-entry hyperlinks among blogs are rather sparse, there are many blogs that are relatively isolated in terms of their observable links. Thus we expect that the extracted communities will only cover a portion of the entire blogosphere. Our ranking-based method starts from highly-ranked blogs to extract dense subgraphs that include popular bloggers.

We utilize the PageRank algorithm [3] to compute the global ranking of blogs. The global ranking score r for all blogs (i.e., nodes in the graph) is given as follows:

$$r_{t+1} = (1-\alpha)P^T r_t + \alpha y, \qquad <6>$$

where r_t is a vector of ranking score r, α is the damping factor, where $y=[1/N]_{N\times 1}$ is a uniform vector where N is the number of nodes, and $P = D^{-1}M$. Recall that M is the mutual awareness matrix, and D is the diagonal matrix with diagonal entry d_{ii} as the row sum of M.

We also use the PageRank algorithm in a different way to extract communities from the graph of blogs. The main ideas of our

clustering method originate from the relationship between clustering and ranking algorithms in terms of the spectral graph theory. In [13], the authors discuss that there is close relationship between the clustering criteria and the properties of a random walk. In [15] the authors apply a spectral ranking to classification problems in graph structures.

Based on [15], we introduce the *association scores* between blogs to indicate how they are well connected within the graph structure. Assume that we have a blog of interest (e.g., the top ranked blog) and want to find a community that is represented by this blog. Then imagine that a random walker always starts from this blog. The random walker will visit a community member more probably than a non-member. In this way, the association score is defined between a blog and the blog where the random walker starts. Based on this intuition, our method starts from a set of blogs each of which represents a community. We refer to these blogs as the seed blogs. By comparing scores of a blog associated with the seed blogs, we decide which community the blog belongs to. Assume that we have m seeds (i.e., they represent m communities). The ranking scores corresponding to the m seeds are computed as follows:

$$R_{t+1} = (1-\alpha)P^T R_t + \alpha Y \ , \qquad <7>$$

where Y is an $N \times m$ matrix and y_{ij} is defined as:

$$y_{ij} = \begin{cases} 1, \text{ if node } i \text{ is the seed of the } j\text{-th community, and} \\ 0, \text{ otherwise.} \end{cases} \qquad <8>$$

This converges to a matrix R = $[r_{ij}]$, where r_{ij} is the association score of blog i to the j-th community. The blog i is assigned to the cluster with the maximum r_{ij}.

Since [15] is to solve classification problem, labeled nodes are available as the seeds to compute ranking. However, in our case, we need to choose a set of seeds from the data set. One way is to use the top m blogs in the global ranking. This is not appropriate since they may not represent different communities. For instance, assume that the first blog and the second blog are mutually aware. Then we want them to belong to the same community instead of creating two different ones.

We address this seed selection issue by choosing seeds iteratively through the ranking-based clustering process. Assume that we have chosen k seeds. Then we want to choose another seed which is not very close to (associated with) the existing seeds. We introduce a boundary of the size of a community (recall that we are not interested in partitioning all the blogs but discovering communities) and exclude blogs in the existing k communities from the candidate of the next seed.

Given k seeds, a set of blogs B is divided to sets C_i:

$$B = \sum_{i=1}^{k} C_i \qquad <9>$$

Then let

$$B_k^N = \sum_{i=1}^{k} C_i^N \qquad <10>$$

where C_i^N is a cluster whose size is bound to N derived by excluding blogs with lower association scores in C_i. The $(k+1)^{th}$ seed is a blog with the maximum global ranking in $\{B \setminus B_k^N\}$.

The process iterates until either (1) a number of clusters have been detected, given the number K, or (2) all the detected clusters are within appropriate size, given the maximum cluster size N_{max}.

6. PERFORMANCE METRICS

We use four metrics—coverage, conductance, interest coefficient, and sustainability—to measure the quality of communities extracted by different methods. The first two are traditional graph theory metrics [2], and the last two are introduced by us in this paper, as we believe that they capture unique blog characteristics.

In the following metric definitions, we use the following notation. $G=(V, E)$ represents the blog graph, where each node v in V represents a blog and each edge (u,v) in E is represents an ordered pair of blogs. For a pair of nodes u and v in V, the edge weight is represented by w_{uv}. We use C_k to denote the members (i.e., the blogs) that belong to the k-th community. The output of a community extraction algorithm is $C= \{C_1 ,..., C_k\}$, resulting in a set of k communities.

Coverage is defined as follows:

$$P(C) = \frac{\sum_{i=1}^{k} \sum_{u \in C_i, v \in C_i} w_{uv}}{\sum_{u,v \in V} w_{uv}} \qquad <11>$$

Coverage measures the fraction of edges that are intra-community. Communities with higher coverage have higher quality. This is intuitive because a larger value of coverage implies more interaction is within communities instead of between community members and non-members.

Conductance is defined for each individual cluster as follows:

$$\Phi(C_i) = \frac{\sum_{u \in C_i, v \notin C_i} w_{uv}}{\min\left(\sum_{u \in C_i, v \in V} w_{uv}, \sum_{u \notin C_i, v \in V} w_{uv} \right)} \qquad <12>$$

A community with a small conductance has higher quality because the number of links pointing to non-members is small relative to the density of either community members or non-community members. We further define the average conductance for a set community as

$$\Phi(C) = \frac{1}{k} \sum_{i=1}^{k} \Phi(C_i) \qquad <13>$$

Interest Coefficient is used to measure how much a community member is interested in his or her assigned community. Intuitively, individual bloggers are supposed to spend more time within their own community. If the majority of a blogger's actions is with members of her own community, then this is a good indication that she is interested (or involved) in this community. A high interest coefficient of a blog community as a whole, suggests that the community members are highly involved. The interest coefficient of an individual blogger m toward an assigned cluster C_k, $I_m(C_k)$, is computed as follows:

$$I_m(C_k) = \frac{\sum_{j \in C_k} w_{mj}}{\sum_{j \in V} w_{mj}} \qquad <14>$$

An aggregated interest coefficient of a cluster C_k, $I(C_k)$ is the average of the interest coefficient of its members:

$$I(C_k) = \frac{\sum_{m \in C_k} I_m(C_k)}{|C_k|}, \quad <15>$$

For a set of clusters C, we compute the weighted average interest coefficient of C by

$$I(C) = \frac{\sum_k |C_k| I(C_k)}{\sum_k |C_k|}, \quad <16>$$

This implies that the larger communities contribute more to the overall interest coefficient of the blogosphere.

Sustainability The three performance metrics defined above are based on aggregated *static* data. We believe that it is important to model analyze temporal dynamics of the blog. A real community in the blogosphere should exhibit cohesiveness. That is, a group of bloggers in a tight community should have sustained interactions over time. We therefore define a new metric, the *sustainability S*, to measure how the membership of communities is sustained over time.

For a community C_i extracted at a specific time t, we define the sustainability S of C_i over a period of time Δt to be the fraction of community members that remain in the community after time Δt has passed. That is

$$S(C_i, \Delta t) = \frac{1}{|C_i|} \max_j (|C_i \cap C_j|), \quad C_j \subset C(t + \Delta t) \quad <17>$$

Then we define the average sustainability for a set C of k extracted communities as follows:

$$S(C, \Delta t) = \frac{1}{k} \sum_{i=1}^{k} S(C_i, \Delta t) \quad <18>$$

7. EXPERIMENTAL RESULTS

We have done extensive experiments on two different independent blog datasets to evaluate our community extraction. The first dataset is obtained from a technology-focused crawler developed at NEC Laboratories America. The second dataset is the WWW2006 Blog Workshop Benchmark data.

7.1 NEC Blog Dataset

In this subsection, we describe the NEC blog dataset, discuss our community extraction results, analyze the output using our performance metrics, examine the temporal properties of the communities, and finally show some interesting communities.

7.1.1 Dataset Description

At the NEC Laboratories America, we have built a focused blog crawler centered on the topic of technology. Due to space limit, we only give a high-level description of the crawler. There are two databases used by the crawler. The first database contains a set of "seed blogs", which initially consist of some well-known blogs with technology focuses. For the seed blogs, the crawler continuously aggregates the RSS feeds and their corresponding entries. For each newly crawled entry, its content is analyzed and the hyperlinks embedded in the content are extracted. If an extracted hyperlink points to another entry and that entry belongs to a blog who is not a member of the seed blogs, then that entry and its blog are stored into the second database. The second database is checked regularly to see if any blog in the database meets the criteria to become a new seed blog (the criteria are based on the number of citations and trackbacks from current seed blogs) and if so, that blog is moved to the first database and starts to be crawled continuously.

The data set we used for the experiments consists of 127,467 entries crawled between July 10th and December 31th in 2005, for a period of 25 consecutive weeks. These entries belong to 584 seed blogs, which are the complete set of seed blogs in our database at the beginning of the first week. In addition, there are totally 40,877 links in the data set. In other words, on average there is only one link for every three entries. In addition, we were able to extract 2,898 trackback links. Figure 5 shows the number of entries in each week in the NEC dataset.

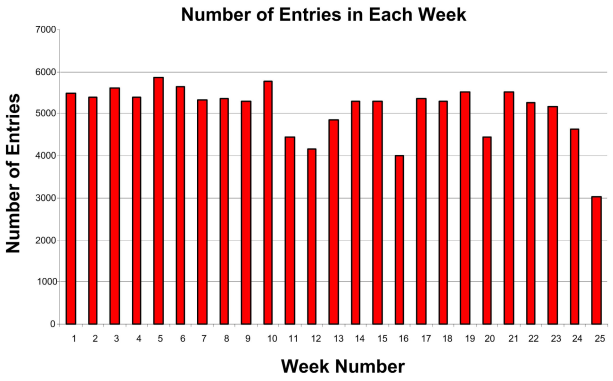


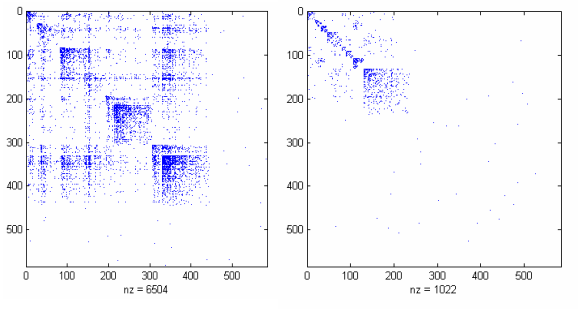
Figure 5: Statistics of the NEC dataset

Entry/Trackback Extraction: A blog is a pair of RSS feed and a web page (i.e., the homepage of the blog) that refers to each other. Entries and trackbacks are extracted from web pages referred to by RSS items in the feed. For each blog, the system generates a set of extraction patterns based on heuristic rules. An entry is extracted as a text region including hyperlinks. If a hyperlink in the region refers to another entry, it is regarded as an entry-to-entry link. When a hyperlink on an entry page e1 is extracted as a trackback and its destination is an entry page e2 of another blog, we regard this as a trackback link from e2 to e1.

7.1.2 Community Extraction

In this section, we extract communities from the aggregated data. That is, blog graphs are built by aggregating all available data in all the 25 weeks. For performance study, we compare the communities extracted by using two different features—the *baseline adjacency matrix* and the *mutual awareness matrix.* When using the baseline adjacency matrix, for an edge *(u,v)* in the blog graph, the total number of entry-to-entry links pointing from blog *u* to blog *v* is used as the edge weight w_{uv}. When using the mutual awareness matrix, w_{uv} represents the mutual awareness score defined as Eq.<4> by considering both entry-to-entry link and trackback links. For each feature, we shall use the ranking-based algorithm describe in Section 5.2 to extract the top 8 communities.

Figure 6 illustrates the communities extracted by using the baseline adjacency matrix as the input feature (for convenience, we refer to these communities as the *baseline communities*) and those extracted by using the mutual awareness as the input feature (we refer to these communities as the *MA communities*). Figure 6(a) is the baseline adjacency matrix with blogs arranged according to their membership in the baseline communities. (A dot in the figure indicates an edge with a non-zero edge weight.)



(a) (b)

Figure 6: Communities extracted using (a) the baseline and (b) the mutual awareness features

Figure 6(b) is the mutual awareness matrix with blogs arranged according to their membership in the MA communities. As can be seen from the figure, in the baseline adjacency matrix, there are more edges with non-zero edge weights. As a result, the baseline communities cover more blogs (more than 400). In comparison, in the mutual awareness matrix, there are fewer edges with non-zero edge weights because of the restriction for mutual awareness. As a result, the MA communities only cover around 250 blogs. However, as we will demonstrate shortly, such smaller communities actually exhibit more intense intra-community interactions.

7.1.3 Performance on the Aggregated Blog Graph

We first use the three static metrics—conductance, interest coefficient, and coverage—to study the performance of the two features (baseline vs. mutual awareness). In the absence of ground truth communities, it is difficult to compare the features – note that each feature results in different edge weights. We address this issue in the following manner. For communities extracted from each feature, we compute the performance metrics on both the baseline adjacency matrix and the mutual awareness matrix. Figure 7 illustrates our experimental design.

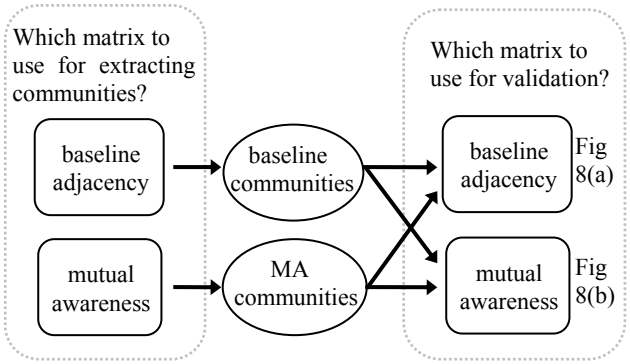


Figure 7: Experiment design showing that for each performance metric, there are four scores. The results are shown in Figure 8.

Figure 8(a) gives the performance metrics computed using the baseline adjacency matrix. As can be seen from the figure, judging from these metrics, the MA communities have better (lower) conductance than the baseline communities. This suggests that members of the MA communities tend not to talk to blogs outside their own communities. The MA communities also have higher interest coefficient, which suggests that members of the MA communities tend to talk to blogs within their own communities. However, the coverage of the MA communities is a little lower than that of the baseline communities. This result is reasonable because the MA communities are much smaller in size—they cover only around 250 blogs compared with the 400 blogs covered by baseline communities. Recall that the coverage measures the fraction of links that are among community members. It is expected that when we use coverage to measure communities that only partially cover the whole blog graph, such a metric usually favors larger community sizes. Figure 8(b) gives the performance metrics computed using the mutual awareness matrix. From the results we can see that in term of low conductance, high interest coefficient, and high coverage, the MA communities clearly outperform the baseline communities. Based on these performance comparisons, we believe that in terms of community extraction in the blogosphere, the mutual awareness is a better feature than simple counts of entry-to-entry links.

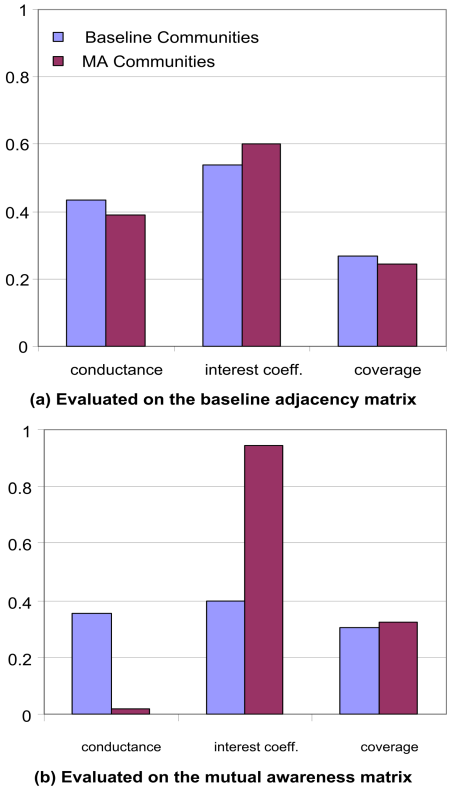


Figure 8: Performance comparison between the Baseline communities and the MA communities using metrics: (1) conductance, (2) interest coefficient, and (3) coverage.

7.1.4 Temporal Properties of the Communities

The sustainability metric computes the change in community membership over time. We start at the 5th week, and then compute changes to the community membership every week (Δt = 1, 2,..) based on the 5th week communities, hence the sustainability at the 5th week is assumed to be 1.

Figure 9 shows the sustainability of the communities obtained from the two features. The communities extracted by using mutual awareness have higher sustainability during most of the subsequent weeks.

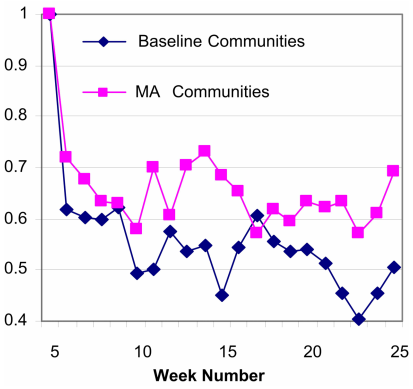


Figure 9: Sustainability comparison between Baseline communities and MA communities

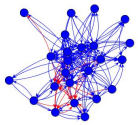
7.1.5 Some Interesting Communities

In this experiment, we use the content of the blog entries to validate our discovered communities. Note that our algorithm purely depends on the structural (hyperlink) information, not on content of blogs. However, intuitively, a valid community should have coherent topics discussed and consistent vocabulary used among community members. Therefore in this experiment, we extract top keywords from each community to see if they form a coherent set. Because different keywords have different language frequency, instead of the absolute frequency of keywords, we compare their deviation from normal language frequency. To do that, for each keyword w, we first compute its relative frequency f_w among contents of *all* blogs; then for a given community, we compute the w's relative frequency f'_w in the content of the blogs in the community. We define the deviation of frequency for w in the given community as f'_w / f_w. For each community, we pick the top keywords with the highest deviations.

Figure 10 shows the extracted keywords together with the subgraph for some of the MA communities. Please notice that to improve readability, we have skipped the stemming step in data pre-processing. From the graph, we can see that among a community, there often exist entry-to-entry links (blue) and trackback links (red) in both directions between a pair of community members, because the mutual awareness scores for such pairs of members are higher. From the list of keywords, we can clearly detect the central topics in the communities, as we suggested on the top of each community. One interesting observation is that our original technology focus has obviously deviated to different areas. Because of space limitation, Figure 10 only shows the results for 4 communities.

A closer look at the Technology Community (Figure 10): This is a very interesting community not only because the blog content within this community exhibits high textual coherence but also many members of this community are CEOs of technology companies—it is a community of *authorities* on technologies concerning communication, wireless, blog, etc.

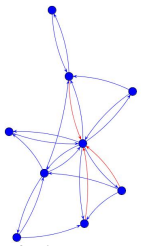
Politics Community:



Number of nodes: 24
Conductance: 0.003
Interest coefficient: 0.98
Coverage: 0.06

joyner, constructor, atta, zarqawi, qaeda, msm, ap, sheehan, michelle, marine, danger, nominee, nomination, democrat, manually, insurgents, terrorists, alito, saddam, sunni, miers, iraqis, hussein, sen, iraqi, cindy, fbi, linked, corps, immigration, captain, baghdad, abu, democrats, desire, pentagon, combat, indictment, terrorist, soldier, army, kick, confirmation, refer, soldiers, terror, palestinian, louisiana, arguments, roberts, liberals, kerry, partisan, plame, insurgency, objects, laden, cia, enemy, iraq, clinton, object, troops, hearings, harriet, afghanistan, senators, coalition, attacks, senate, syria, guard, islamic, military, conservatives, gop, supreme, ping, surveillance, killed, propaganda, libby, liberal, scandal, judicial, weapons, muslim, spokesman, republicans, suicide, bombings, conspiracy, conservative, james, muslims, senator, bin, intelligence, commission, enforcement

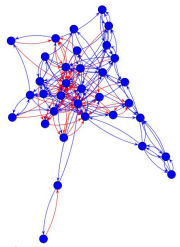
International Community:



Number of nodes: 9
Conductance: 0.005
Interest coefficient: 0.86
Coverage: 0.006

shanghaiist, shanghai, uzbekistan, hong, kong, beijing, taiwan, uranium, china, chinese, niger, russia, simon, asia, japan, indian, pakistan, india, bird, japanese, tokyo, russian, flu, asian, gop, singapore, minister, foreign, village, democracy, thread, central, daily, ministry, authorities, korea, alleged, opposition, ties, relations, afghanistan, iran, farmers, prime, cent, nuclear, region, visited, economic, english, ambassador, husband, corruption, countries, deputy, western, british, korean, borders, germany, rural, restaurant, international, republic, dogs, eu, festival, accused, claim, regime, rice, africa, arms, domestic, residents, government, cultural, official, german, economist, north, cia, population, rising, trip, rule, wilson, protest, parliament, democratic, regional, documents, joseph, east, pacific, river, officials, apos, country, propaganda

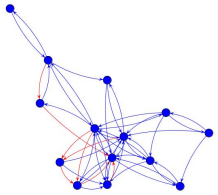
Technology Community:



Number of nodes: 34
Conductance: 0.05
Interest coefficient: 0.86
Coverage: 0.004

skype, optional, outlook, jeremy, ibm, dell, technorati, marc, bloglines, tb, mtcmtauth, mtcmthome, mtcmtmail, collaboration, undefined, tasks, scoble, communication, immediate, tag, customer, productivity, tags, typekey, wiki, vendors, suggestions, xp, checked, marketing, signing, tagging, podcasting, alan, brands, seth, podcasts, podcast, vista, feedburner, conversation, identity, alex, weblogsinc, innovation, winer, conversations, architecture, blogging, social, mary, communities, blogosphere, messaging, discussed, feedback, myspace, recorded, marketers, format, aggregator, integration, formats, rss, schedule, sun, employee, minute, ui, adam, friend, windows, gmail, enterprise, entries, adjab, open, discussion, journalism, supports, competitors, clients, apps, jeff, message, session, adoption, email, pcs, client, notes, beta, icio, developer, folks, relations, external, blog, sign, tools

Economics Community:



Number of nodes: 14
Conductance: 0.04
Interest coefficient: 0.78
Coverage: 0.001

ebffed, loan, economics, liberty, insurance, economist, typepad, pub, payment, drugs, reference, drug, listed, essay, wal, moral, health, taxes, welfare, mart, amendment, property, affairs, profits, eu, guardian, wealth, economic, poverty, immigration, theory, prices, gasoline, suppose, rate, carnival, medical, sector, private, goods, roundup, argued, agent, odd, britain, politicians, nuclear, costs, russian, trade, rates, income, doctors, minimum, consequences, treatment, profit, belief, jose, argument, housing, markets, higher, estate, productivity, care, england, ought, workers, faith, captain, constitution, weblogs, british, disagree, principle, balance, institute, fees, benefits, supply, aid, warming, policy, reduce, policies, correct, logic, institutions, afford, government, savings, senator, relative, insurgency, anybody, union, laws, oil, drink

Figure 10: Four resulting MA communities demonstrating topical cohesiveness

7.2 Workshop Dataset

Similar to the previous subsection, we describe the WWW2006 Blog Workshop dataset, discuss our community extraction results, analyze the output using our performance metrics, examine the temporal properties of the communities, and finally show some interesting communities.

7.2.1 Dataset Description

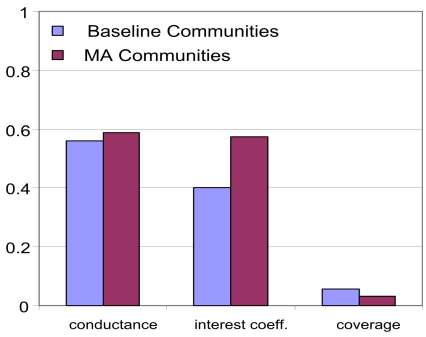
We apply our community extraction algorithm on the dataset provided by the workshop organizer. Compared with the dataset used in the previous section, this workshop dataset is of much larger scale. It contains 8.37 million entries from 1.43 million different blog sites during the time between July 4th and July 24th, 2005. Because our community extraction algorithm is link-based, instead of the whole dataset, we studied the subset of blogs that contain at least one link. By this restriction, we are able to narrow down the number of blogs to around 122K. Then we build a blog graph with 122K nodes and use the numbers of links among blogs as edge weights. For computing mutual awareness, we use two types of links, blog-to-blog links and Web-page-co-citation links, both of which are derived from the original dataset. (However, at the time of this writing, our experiments on data with Web-page-co-citation links are still ongoing. Therefore, for all the MA results reported here for the workshop dataset, the mutual awareness is computed based only on blog-to-blog links.)

With a dataset of such a large scale, there could easily be hundreds of communities. In the preliminary study, we focus on the first 24 communities that are discovered by our algorithm.

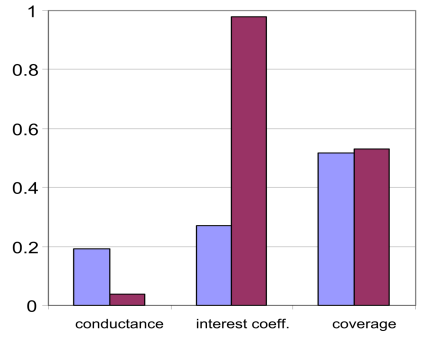
7.2.2 Performance on the Aggregated Blog Graph

We again use the three static metrics—conductance, interest coefficient, and coverage—to study the performance of the two features (baseline vs. mutual awareness). This study is based on blog graphs obtained from aggregating all available data in 21 days. For the same reason as before, we compute the performance metrics on both the baseline adjacency matrix and the mutual awareness matrix.

Figure 11(a) shows the performance comparison based on the baseline adjacency matrix. As can be seen, judging by the baseline adjacency matrix, the MA communities have somewhat worse (higher) conductance and worse (lower) coverage. However, in this case, the MA communities have better (higher) interest coefficient. This suggests that from individual community member point of view, the MA communities are better. This is because on average, an individual community member tends to interact more with members within the same community as he or she belongs to. Figure 11(b) gives the performance metrics computed using the MA adjacency matrix. Again, judging by the mutual awareness scores, in term of low conductance, high interest coefficient, and high coverage, the MA communities outperform the baseline communities.



(a) Evaluated on the baseline adjacency matrix



(b) Evaluated on the mutual awareness matrix

Figure 11: Performance comparison between the Baseline communities and the MA communities using metrics (1) conductance, (2) interest coefficient, and (3) coverage.

7.2.3 Temporal Properties of Communities

To compute the sustainability metric, we start at the end of the first week (July 10th) and compute the changes to the community membership at the end of the second week (July 17th) and the end of the third week (July 24th). Figure 12 shows the sustainability of the communities obtained from the two features. As can be seen, the the MA communities have better sustainability than the baseline communities in both the second and the third weeks.

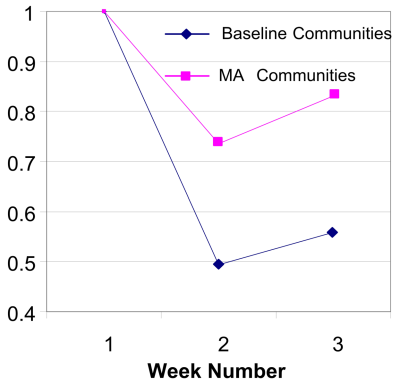


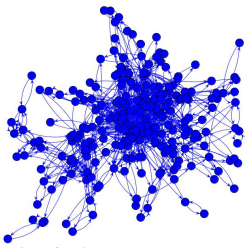
Figure 12: Sustainability comparison between baseline communities and MA communities

7.2.4 Interesting Communities

In this part, we show several interesting communities that our algorithm was able to discover. For some of these communities, we were able to detect the cohesive topics easily from the top keywords; for others, we were not able to find central topics. We examine one example for each case.

Communities with obvious central topics: In Figure 13 we first show some interesting communities from the top keywords of which we were able to detect topics (with topic name suggested).

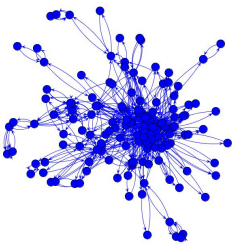
Mystery Community ("harry potter" "buffy the vampire slayer")**:**



Number of nodes: 239
Conductance: 0.0004
Interest coefficient: 0.99
Coverage: 0.02

remus, sirius, severus, draco, fic, snape, wand, fandom, hermione, malfoy, weasley, jkr, hbp, lupin, voldemort, spoilers, hogwarts, spoiler, ron, slash, lily, dumbledore, ginny, harry, canon, summary, buffy, wan, arse, reed, chapter, gaze, pale, james, characters, glanced, drew, orlando, grinned, xander, rowling, angst, sighed, alexander, plot, nodded, rating, willow, requested, fred, fictional, peter, vampire, charm, elf, jaw, ministry, wizard, meme, shrugged, curse, beta, trio, whispered, arthur, chapters, jedi, scar, veronica, novel, swallowed, stroking, fingers, grin, stared, bella, quietly, victor, robe, paused, novels, hp, magical, utterly, tagged, roy, pulls, scent, moody, amused, shook, softly, author, cheek, forehead, reflection, lips, spoiled, beneath, shaking

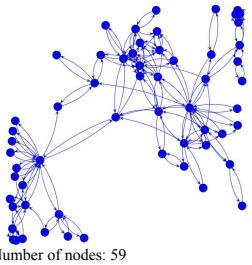
Politics Community ("valerie plame leak investigation")**:**



Number of nodes: 146
Conductance: 0.1
Interest coefficient: 0.96
Coverage: 0.02

fitzgerald, plame, novak, rove, mcclellan, covert, gop, cheney, operative, niger, memo, valerie, karl, cia, leaked, cooper, classified, leak, republicans, ais, republican, administration, wilson, investigation, patriot, reporters, shakespeare, democratic, evolution, ambassador, democrats, outing, scandal, bush, deputy, alternate, deserves, nominee, reporter, senate, governor, jury, miller, senator, mercury, corruption, roe, iraq, presidential, nomination, elections, congress, journalists, roberts, conservatives, conspiracy, president, judy, chairman, reform, officials, committee, hussein, saddam, fbi, iraqi, rising, intelligence, constitutional, troops, judicial, identity, voters, ongoing, ohio, involvement, senators, campaign, political, election, publicly, supreme, economy, fran, progressive, ethics, journalist

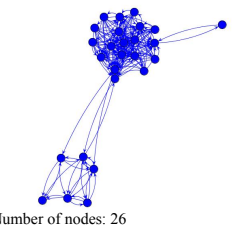
Middle-East Community:



Number of nodes: 59
Conductance: 0.18
Interest coefficient: 0.96
Coverage: 0.01

israel, palestinian, jews, israeli, jewish, sharon, islamic, muslims, bomber, civilians, musings, saudi, islam, journalism, editorial, muslim, arab, terror, bombers, religious, gates, protest, terrorism, terrorists, leslie, residents, minister, moderate, soldiers, leadership, hussein, saddam, buses, terrorist, journalist, ruling, qaeda, occupation, radical, attacks, podcast, kris, bombing, communities, authority, hitler, territory, journalists, prime, tagged, bombings, bloggers, politics, reporters, suicide, principles, londoners, david, nations, policies, opposition, divine, tel, ban, silent, britain, statements, allah, egypt, democratic, principle, behalf, policy, religion, demands, column, iraq, democracy, wounded, conflict, minority, zealand, writes, representative, afghanistan, declaration, hatred, election, affairs, reporter, mainstream,

Technology Community:



Number of nodes: 26
Conductance: 0.0002
Interest coefficient: 0.99
Coverage: 0.007

linking, weblogs, ads, blogs, shiny, photoshop, filed, squad, google, email, cellphone, podcasting, indie, insider, device, weblog, flash, wireless, wonders, programming, marc, usb, erik, hack, olsen, producing, nextel, comments, hd, intel, sandals, bloggers, explains, devices, slot, blogging, thompson, podcast, tech, bid, cbs, casting, sci, sprint, hybrid, outlook, firefox, verizon, reed, vince, nbc, cingular, releases, networks, interface, bag, foster, manufacturers, podcasts, automotive, zoom, samsung, gps, reports, solar, broadband, users, integrated, shark, portable, designer, battery, launch, toyota, upgrade, finds, stewart, powered, layer, cable, ebay, vehicles, sirius, receiver, posts, applications, technologies, digital, developer, priced, platform, disk, industry, maker, victor

Figure 13: Four resulting MA communities demonstrating topical cohesiveness

A closer look at the Technology Community (Figure 13): This is not really a community of people. Many of them are blogs providing shopping, TV, or cinema guides, or promoting products such as electronic gadgets, etc. (but they are not spammers). From the diagram, it looks like there is a very cohesive cluster together with a relative looser cluster; however, the weights of the edges between two sub-clusters are indeed high enough to integrate the two into one community.

Community with no obvious central topic: The community in Figure 14 is one of the communities that we were not able to discover cohesive topics from the top keywords. However, from the metrics of conductance and interest coefficient we can see that members in the community almost exclusively interact with other members in this same community and seldom communicate with non-members.

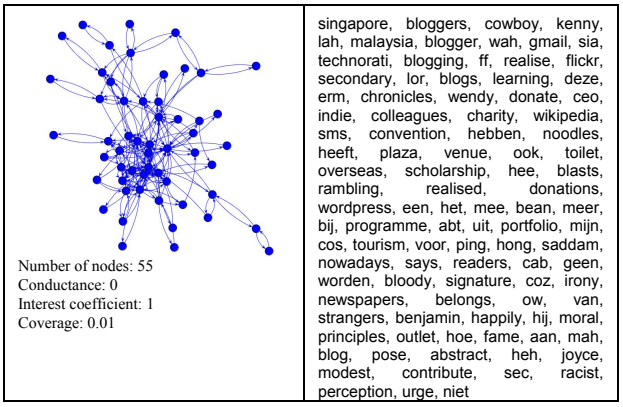


Figure 14: An MA community that has no obvious central topic

A closer look at the "No topic" Community (Figure 14): The members of this community are real online friends—most of them are 20-30 year-old youngsters living in Sigapore and Malaysia—they are aware of each other and frequently interact with each other but the topics of their talks are not restricted to a specific area. This is a closed community of bloggers from other countries, which could be a reason for extremely low (0) conductance.

7.2.5 Mutual awareness data available in the NEC dataset vs. the Workshop dataset

Due to lack of trackback information in the workshop dataset, the computation of mutual awareness for extraction communities depends solely on bi-directional entry-to-entry links. In NEC dataset which partial trackback information is available, we observe high coincidence between trackbacks and bi-directional entry-to-entry links. It is worth to note that for some communication which only uni-directional entrylinks are observed, trackbacks can be regarded as strong evidence of mutual awareness and can be used to assure or supplement the strength of the communication between two blogs, as we have shown in our experiment on the NEC dataset. Since the workshop dataset collects blog data within only three weeks, data concerning two blogs with bi-directional entry-to-entry links are rather sparse—it would be useful to have other action information, such as trackbacks, to infer mutual awareness.

8. CONCLUSION

We presented a novel framework for the discovery of blog communities based on the following two key insights: (a) communities form due to individual blogger actions, and these actions must be *mutually observable*; (b) semantics of hyperlink structures are different from the traditional web analysis problem.

Our computational approach to community extraction has two steps – (1) computational models of mutual awareness, and (2) a ranking-based clustering method. Mutual awareness coefficient between two bloggers is a function of the type of action, the number of actions for each type, and the time of the action. We use an iterative, ranking-based clustering scheme on the mutual awareness matrix to determine the communities.

We conducted extensive experiments on two independent datasets (NEC and Workshop), with excellent results. Four validation metrics are defined and used to analyze the results – (a) coverage, (b) conductance, (c) interest coefficient and (d) sustainability. Our findings indicated that the mutual awareness matrix outperforms the traditional adjacency matrix on most of the performance measures, over both datasets. Even in the absence of ground truth, these results are strongly suggestive of the value of our approach.

There remains some open issues to explore for further research: (a) An explanation for the rises and falls on the sustainability curves would require sophisticated analysis on, for examples, the changes of number of links and entries, possible events that cause new communication established between two blogs, etc. Besides, assuming constant number of communities to compute the sustainability is unnatural since it is possible that old communities die and new communities born over time. It would be most interesting to further examine the evolution of these communities. (b) Uni-directional entry-to-entry links could also lead to mutual awareness even if they are not reciprocated. Many bloggers regularly perform vanity searches or even subscribe to an RSS feed of all links to their blog and thus are acutely aware of when other bloggers link to their blog. The RSS feed suggests some useful information to extend the inference of mutual awareness. (c) The performance metrics as constructed are not the best ones for comparing mutual awareness versus baseline communities. It would be preferable to include metrics that use something other than links, since community extraction is so heavily dependent on linking behavior. For example, an interesting metric could incorporate the textual coherence explicitly (as opposed to implicitly via keyword extraction).

Our initial results point to several promising research directions, including:

(1) Community sustainability analysis: discovering the evolution of a community based on the idea of sustainability.

(2) Topical blog community extraction: extracting topical communities based on link semantics (mutual awareness) and content analysis (keyword extraction).

(3) Analysis of roles in blog communities: recognizing and identifying specific roles in blog communities based on a Bayesian representation of individual blogger actions and interactions. The Bayesian representation can also generate a large scale synthetic datasets, which can useful for establishing ground truth.

9. REFERENCES

[1] A. BLANCHARD (2004) *Blogs as Virtual Communities: Identifying a Sense of Community* http://blog.lib.umn.edu/blogosphere/blogs_as_virtual.html.

[2] U. BRANDES, M. GAERTLER and D. WAGNER (2003). *Experiments on Graph Clustering Algorithms*, 11th Annula European Symposium on Algorithms - ESA 2003, Springer, 568-579, Sep. 2003, Budapest, Hungary.

[3] S. BRIN and L. PAGE (1998). *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems **30**(1--7): 107--117.

[4] I. DHILLON, Y. GUAN and B. KULIS (2004). *A Unified View of Kernel k-means, Spectral Clustering and Graph partitioning*, July, 2004.

[5] P. DOURISH (2001). Where the action is : the foundations of embodied interaction. MIT Press Cambridge, MA.

[6] G. W. FLAKE, S. LAWRENCE and C.L.GILES (2000). *Efficient identification of web communities*, In Proc. 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining,

[7] T. H. HAVELIWALA (2002). *Topic-sensitive PageRank*, Proceedings International WWW Conference, 517-526,

[8] H. INO, M. KUDO and A. NAKAMURA (2005). *Partitioning of Web Graphs by Community Topology*, Proceedings of the 14th international conference on World Wide Web, 661-669, Chiba, Japan.

[9] Q. JONES (1997). *Virtual-Communities, Virtual Settlements & Cyber-Archaeology: A Theoretical Outline.* JCMC **3**(3).

[10] J. M. KLEINBERG (1999). *Authoritative sources in a hyperlinked environment.* J. ACM **46**(5): 604-632.

[11] R. KUMAR, J. NOVAK, P. RAGHAVAN and A. TOMKINS (2003). *On the bursty evolution of Blogspace*, Proc. Of the 12th International Conference on World Wide Web, Budapest, Hungary.

[12] H. RHEINGOLD (2000). The Virtual Community: Homesteading on the Electronic Frontier. The MIT Press.

[13] J. SHI and J. MALIK (2000). *Normalized Cuts and Image Segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8): 888-905.

[14] B. L. TSENG, J. TATEMURA and Y. WU *Tomographic Clustering To Visualize Blog Communities as Mountain Views*, 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics,

[15] D. ZHOU, J. WESTON, A. GRETTON, O. BOUSQUET and B. SCHOLKOPF (2004). *Ranking on Data Manifolds*, NIPS 2003.