On the Structure, Properties and Utility of Internal Corporate Blogs*

Pranam Kolari‡ kolari1@umbc.edu
Tim Finin‡ finin@umbc.edu
Kelly Lyons◇ klyons@ca.ibm.com
Yelena Yesha‡ yeyesha@umbc.edu
Yaacov Yesha‡ yayesha@umbc.edu
Stephen Perelgut◇ perelgut@ca.ibm.com
Jen Hawkins◇ jlhawkin@ca.ibm.com

‡Department of CSEE, University of Maryland, Baltimore County, Baltimore, MD 21250, USA
◇IBM Toronto Laboratory, 8200 Warden Ave, Markham, Ontario, L6G 1C7, Canada

Abstract

Weblogs, or blogs are radically changing the face of communication within enterprises. While at the minimum blogs empower employees to publicly voice opinion and share expertise, collectively they improve collaboration and enable internal business intelligence. Though the power of blogs within organizations is well accepted, their properties, structure and utility has not yet been formally analyzed. In this paper, we study the use of blogs within a large corporation to reveal some of the interesting characteristics. We propose new techniques to model the reach and impact of posts using the corporate hierarchy. We discuss how such a technique can feed into tools that identify the reach of blog posts, and the emergence of trends and experts within an organization.

1. Introduction

The growth of blogs has been phenomenal over the last few years. Unlike e-mail and messaging, it offers a more open medium of communication, enabling authors (bloggers) to reach out beyond their social networks, make new connections, and form communities. Collectively, this makes the community of bloggers highly influential. Tapping into this new channel to listen to and interact with their customers requires new initiatives from corporations. Businesses, both large and small, now recognize the power of blogs for engaging with customers, developing trust around their products and services, and improving media visibility. Most corporations are now blogging publicly, either through product bloggers, evangelists or CEOs. However, this covers only one side of the story.

A second key aspect of blogs for business is their use within the organization. Internal corporate (enterprise or business) blogs encompass all non-public blogs hosted within the organization on their intranets. Employees use such blogs during the course of their daily responsibilities, to share expertise on products and services, to voice opinions, and to initiate discussions on issues of interest to other employees. Blogs protect the ownership of employee ideas. Overall, blogs are viewed as a collaboration tool enhancing productivity, and as an enabler for business and competitive intelligence. They are also considered as a tool for *workforce journalism*, an activity that can influence an organization's external presentation through public facing blogs or other communication channels.

Consequently, the market for internal blogging tools is growing [4]. Blog publishing vendors who until now catered to a general audience are now positioning packaged products that address internal enterprise needs. Improving the utility of such packages involves understanding how existing tools are used within organizations and how they can be (and are) used for internal business intelligence. Though it is widely accepted that blogs enable the emergence of thought leaders and experts, and the identification of popular themes [9], it is still not clear as to what is the best away to achieve this. These questions form the primary motivation for work reported in this paper.

Our key contributions are:

- Our study is the first to comprehensively characterize a community of blogs and the social network it materializes within an enterprise.
- We propose new techniques to model the impact of a blog post based on its reach in an organizational hierarchy.
- Our findings enable development of new tools and techniques that facilitate improved utilization of blogs within organizations.

The rest of the paper is organized as follows. Section 2 briefly describes the dataset used in this work. In section 3 we discuss the growth of internal blogs and the use of tags. We detail the structural properties of the network in section 4. Motivated by the characteristics of internal blogs and the context of its use, we propose a model for evaluating reach in section 5, and discuss its utility for internal business intelligence. Finally, before concluding the paper, we discuss the implications of our work to internal corporate blogs in general.

2. Analyzed Collection

This work is based on internal blogs within IBM between November 2003 and August 2006. IBM is a global technology corporation with over 300000 employees, and 23000 registered blog users. Blogs are published using an extended version of the Roller[1] platform, an Apache powered open source Java implementation also in use by Sun, and other corporations.

Each blog is owned by an employee, or a group of employees, and there are a total of around 23500 blogs. These blogs host 48500 posts

*Partial support was provided by an IBM CAS Fellowship and by NSF awards NSF-ITR-IIS00326460 and NSF-ITR-IDM-0219649.

[1] http://rollerweblogger.org

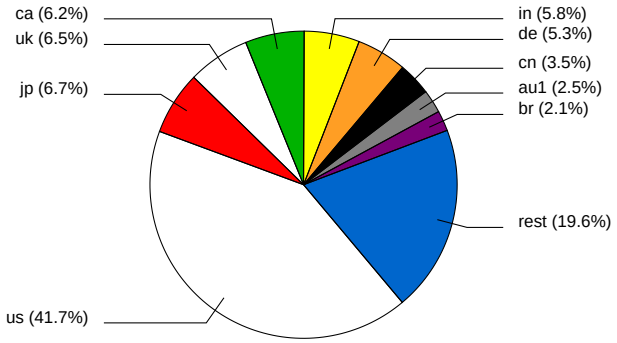

Fig. 1: *Geographical Distribution of bloggers in the collection as denoted by country tags.*

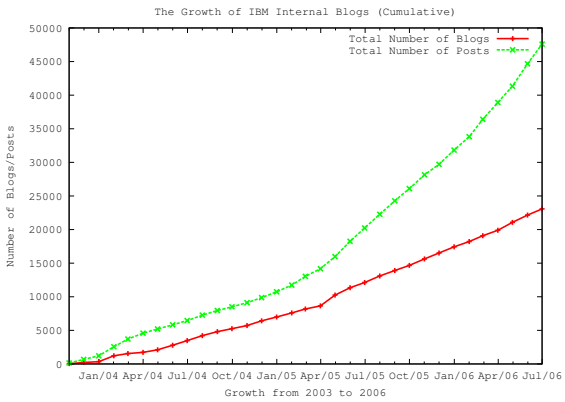

Fig. 2: *Growth of blogs and hosted posts has been phenomenal, with the number of hosted blogs doubling every 10 months.*

with a similar number of comments including trackbacks. Posts carry with them a timestamp, author and tags. Tags associate content to a folksonomy of topics as perceived by the author. The collection of tags consists of around 7200 distinct tags. For every employee owning a blog, information on their geographical location, and to the position and chain in the corporate hierarchy is available. The location-specific distribution of posts in this collection is shown in figure 1. The distribution closely mirrors the use of blogs on the Web [29] i.e. led by English speaking areas, but followed closely by Asia and Europe.

3. Nature of the Internal Blogosphere

We first discuss some of the basic characteristics of blogs, as they relate to growth and their use within the enterprise.

3.1 Growth and Attrition of Users

The external blogosphere has been doubling every six months for the past two years [29]. Their growth internally is not as high, but healthy nonetheless, doubling at a little less than a year. Figure 2 shows the number of blogs and posts on a cumulative scale. The divergence between blogs and posts shows an interesting trend on how the blogging community is better engaging new adopters, and encouraging them to post frequently, therefore retaining them.

To find exactly how the creation of new blogs and posts trend over time we plotted the number of new blogs and posts per month, as shown in figure 3. Two distinct spikes characterize this growth. The first, early in January 2004 was around the time when internal blogs were initiated within the organization. However, the second sharp rise around April or May 2005 was critical to the growth of blogs for two significant reasons, (i) the period following this is characterized by a dramatic increase in blog posts, and (ii) number of new blogs created every month has doubled from 500 to 1000 from before to after, suggesting that adoption was catalyzed. It turns out that at this time the organization officially embraced blogging as a communication medium, and formally specified its policy and guidelines for both internal and external blogs. Evidently, having formal policies and a top-down guidance embracing blogs is key to the adoption of blogs by employees.

To better understand blogger adoption and attrition we computed the retention of users at monthly intervals using the following definition:

Definition *A user who posted on a specific month, is considered retained if he or she reposts at least once in the following x months.*

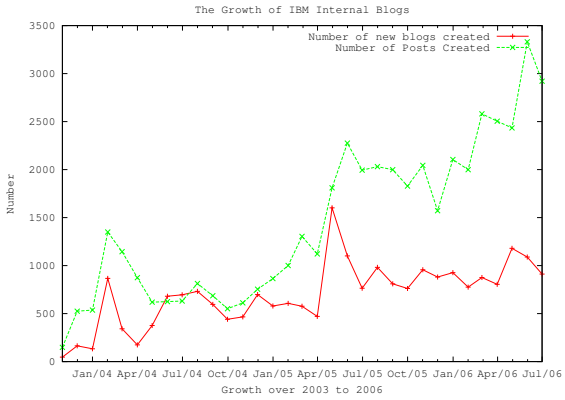

Fig. 3: *The creation of new blogs and posts are also tied to organizations publishing formal policies, which explains the second spike.*

We set the value of x to 6 months. All users not retained are considered lost by attrition. Figure 4 depicts the rates of attrition and retention of users. In line with the previous trends, two distinct spikes characterize this chart, one at the initiation of internal blogs, and the other when blogging policies were formally released. Though not all adopters at these spikes were retained, it did enable the somewhat reluctant bloggers to post more frequently. Overall, the gradual rise of the retention curve over attrition underscores improving ability of the community to retain new adopters.

3.2 Use of Tags

Tagging is fast becoming a common way of associating keywords (tags) to organize content. The collection of tags within a specific system or application defines a folksonomy. If tagging is the means, folksonomy is the result. What drives their popularity is simplicity.

Tags on blogs are no different, and their use has been rising. We analyzed to see how tags, and the concept of folksonomy is being adopted by internal blog authors. As shown in figure 5, close to 80% of all posts are tagged. The chart suggests that tag usage was higher during the early phase of internal blogs. We believe this is because the earliest of adopters were quite adept to the idea of using tags. The addition of new bloggers, and dilution of the contribution

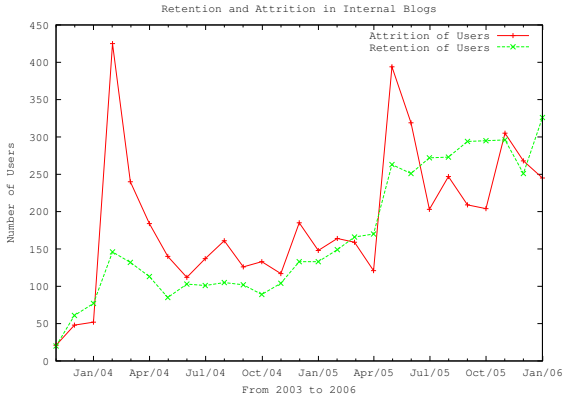

Fig. 4: *The retention of users has gradually taken over attrition showing how the community is reaching critical mass.*

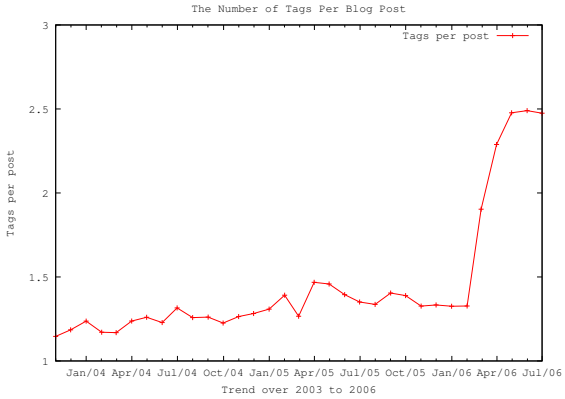

Fig. 6: *The number of tags used per post is increasing. The sharp rise is attributed to an upgrade in the internal blogging that encouraged tagging, and to the addition of "bookmark-it" feature that showed its value across folksonomies.*

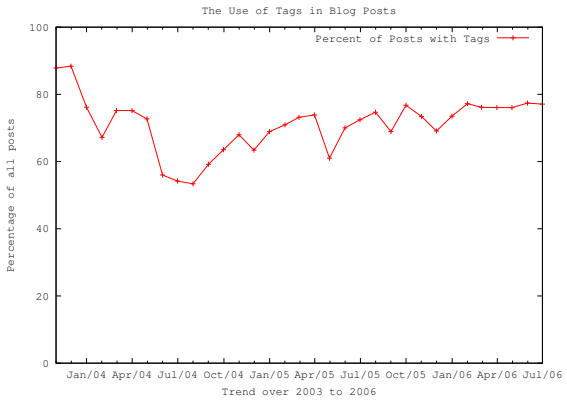

Fig. 5: *Use of Tags in Blog Posts has seen a gradual rise, with the creation of new tools that show their value, and the general exposure of users to folksonomies.*

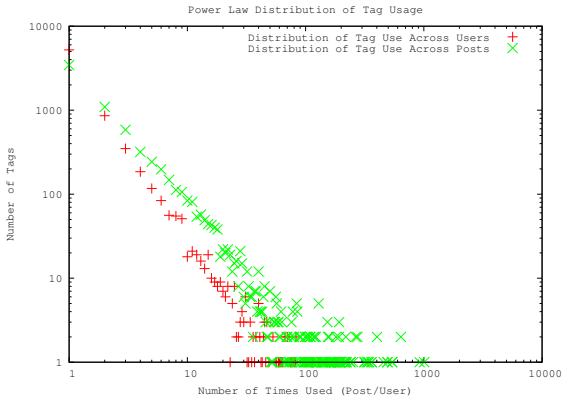

Fig. 7: *The distribution of tags based on their occurrence across all posts, and by the number of authors using them.*

of early adopters to newly made posts explains an early reduction in tag usage. However, as shown in the figure, the use of tags has only grown over the last two years.

The quality of a folksonomy is directly related to properties of tags it hosts. Only now is the concept of *folksonomy quality* receiving required attention [15, 22], and forms key to understanding the utility of tagging within an application. Specifically, when is a folksonomy considered to be *good* or *bad* is still unanswered. We study three attributes that have a potential bearing on quality, (i) the number of tags per post, (ii) the distribution of usage of a specific tag across all posts, and (iii) the distribution of usage of a specific tag across users making these posts.

Figure 6 shows a rise in number of tags per post. This indicates that the descriptive value of tags is improving. The sharp increase in the number of tags per post around January 2006 is attributed to upgrade in the blogging platform, that made adding tags easier, and to the integration of a *bookmark it* feature to blog posts that automatically exported tags to an internal bookmarking tool [24]. From an enterprise standpoint, the interoperability of tags across multiple folksonomies encourages the use of more descriptive tags.

A specific tag provides better value to a folksonomy when used many times. The use of a tag within a folksonomy only once does not reflect well on overall quality, though it could be useful to individual users. Figure 7 shows the distribution of tags based on the number of occurrences in the folksonomy. The x-axis represents the number of times a specific tag is used, and the y-axis gives the number of such tags, on logarithmic scales. The usage follows a power-law distribution indicating that a small number of tags are used with a high frequency, and a large number of them are rarely used. Such a property of tags renders them useful for trend analysis.

A specific tag provides better value to a folksonomy when used by multiple users. A tag being used by only one user reflects on a very narrow scope, with utility only to the blog author. The second plot in figure 7 represents on the x-axis the number of users using a specific tag. More authors using the same tags reflects well on quality[2]. Clearly, a subset of popular tags is used by a large number of users. We believe the relationship between the slopes and offset of lines that fit these plots could have useful implications on folksonomy quality. A more accurate estimation of quality of a folksonomy requires further analysis.

Since tags are less susceptible to spam in a controlled enterprise environment, the high use of tags presents new opportunities for

[2] Semantic disambiguation has to be appropriately incorporated

trend analysis, and towards organizing and navigating blog posts contextually.

3.3 Links from posts

Using posts from 2 months, we analyzed how many posts feature out-links (hyperlinks), and if they do, where do they point to. 60% of all posts featured out-links of one form or the other. Out of these posts, close to 70% had links to the domain of the enterprise, 50% to other domains and 22% to other internal blogs. This leads us to two observations, (i) employees typically blog about themes of interest to the organization they work for, and (ii) since the overall post to comments (including trackbacks) ratio is close to one, trackbacks are a less favored form of conversation threading.

4. Network Characteristics

To study the structural properties of internal blogs, we generate a directed graph $G(V, E)$, where V is a non-empty finite set of vertices or nodes, and E is a finite set of edges between them. Every unique user u, independent of whether they own a single blog or multiple blogs, represents a vertex in G. A directed edge e from node u to node v exists in G, if user u has commented or trackbacked a blog post made by user v. Each such edge represents a *conversation*. We call such a graph, a *blog conversation graph*, since it reflects on conversations across users through blogs. G also represents a social network across all users.

Further processing was made on G to eliminate self-loops, to collapse multiple edges between nodes into a single edge, and to prune disconnected nodes. Almost 75% of the nodes in the graph were completely disconnected. Each such user, on an average had either one post or had just created a blog template without creating blog posts, or making comments on other posts. After processing, the complete graph consisted of 4500 nodes with 17500 edges.

In the rest of this section we discuss some of the structural properties of this network, and its implications to internal blogs. All our experiments make use of the JUNG[3] toolkit.

4.1 Degree Distribution

The degree distribution of a network is significant in understanding the dynamics of a network and its resilience to the deletion of nodes [5, 6]. For every node u in G, the in-degree d_{in} and the out-degree d_{out} is computed as the number of incoming and outgoing edges respectively. The in-degree $P(d_{in})$, and out-degree distributions $P(d_{out})$ is then plotted on a log-log scale, and the power-law exponents γ_{in} and γ_{out} computed using a line fit.

The in-degree and out-degree distribution of G follows a power-law as shown in figure 8 and figure 9, with $\gamma_{in} = -1.6$ and $\gamma_{out} = -1.9$. This is slightly lesser than their values found on the Web ($\gamma_{out} = -2.67, \gamma_{in} = -2.1$) [7], but comparable to e-mail networks ($\gamma_{out} = -2.03, \gamma_{in} = -1.49$) [11]. In the context of blogs, this scale-free property of the network shows the *resilience of the community to user attrition*.

4.2 Degree Correlation

Another interesting property of communication media is degree correlation. In blogs, it measures the reciprocal nature of comments i.e. *Do users who receive a number of comments, make a similar number of comments?* We adopt the approach used previously in call graph networks [25], and plot the average out-degree of all nodes with the same in-degree. Results are shown in figure 10. The correlation holds for smaller degrees, but diverges randomly at higher values, possibly because of insufficient data points at such values. In general, active users in the community host conversations on their own blog, and contribute to conversations on other blogs.

[3] http://jung.sourceforge.net/

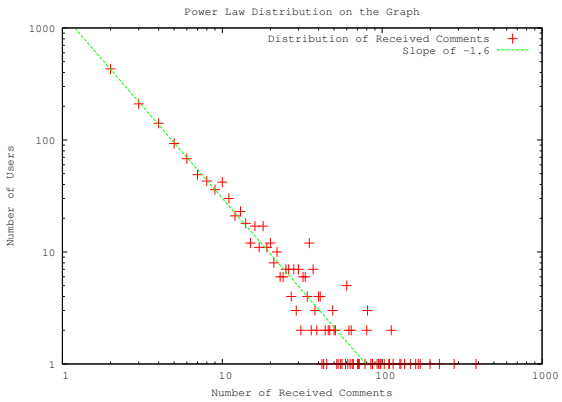

Fig. 8: *The in-degree of the network follows a power-law with slope -1.6. A few users generate most of the conversation.*

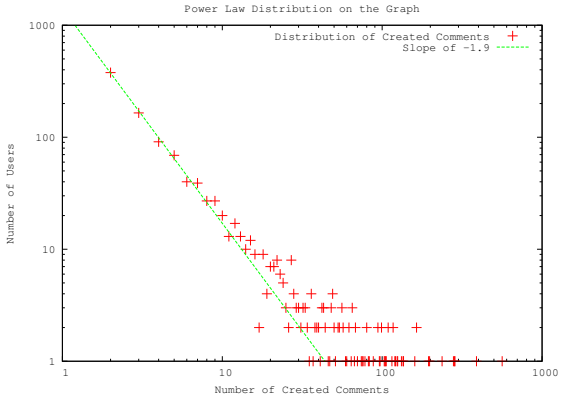

Fig. 9: *The out-degree of the network follows a power-law with slope -1.9. A few users are part of many conversations.*

4.3 Graph Ranking Correlation

The growth of the Web has popularized multiple node ranking approaches that work on a graph. On the Web, these techniques provide the importance of a Web-page. In social networks they give a measure of the popularity or social importance of a node. Three approaches are commonly used to rank nodes in a graph, in-degree, HITS [19] and PageRank [28]. Unlike PageRank and HITS, in-degree is easily computed as the number of incoming edges.

The PageRank of a node u on a graph is computed as:

$$p(u) = \frac{q}{N} + (1-q) \sum_{v:v \to u} p(v)/d_{out}(v)$$

where N is the total number of nodes, q is a constant with $0 < q < 1$ and $(1-q)$ is the dampening factor, $j \to i$ indicates the existence of an edge from node v to node u, and $d_{out}(v)$ is the out-degree of v. The HITS ranking technique computes the hub and authority score. A good hub is one that points to a number of authoritative sources, while a good authority is one that is pointed to by many hubs. The Hub and Authority scores for a node u, represented as $H(u)$ and

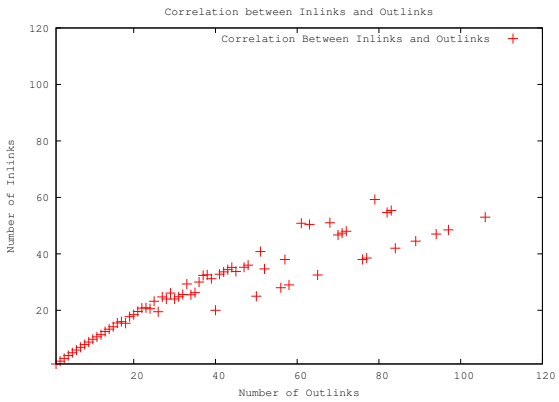

Fig. 10: *A high correlation between in-degree and out-degree shows the highly reciprocal nature of blog comments.*

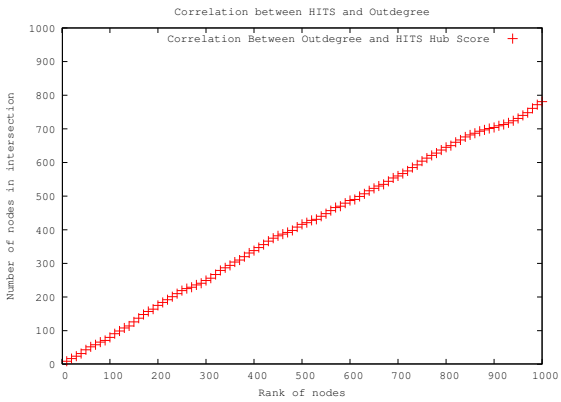

Fig. 12: *The correlation between HITS hub score and out-degree suggests that connectors in the community can be easily identified using their out-degree.*

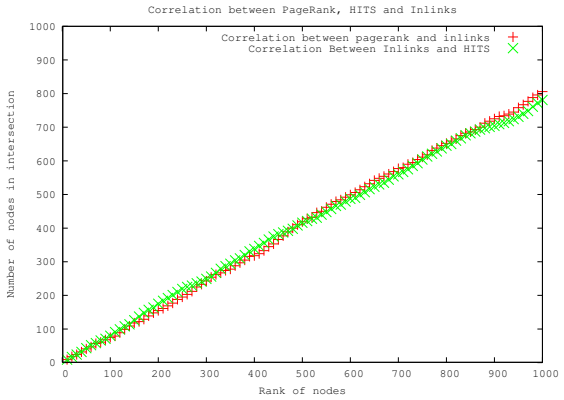

Fig. 11: *The correlation of in-degree with both PageRank and HITS suggests that in-degree can indicate a good approximation of authority.*

C-P	us	jp	uk	ca	in	de	cn	au1	br
us	41.4	0.3	8.9	4.4	0.6	1.4	0.2	1.2	0.4
jp	2.1	4.3	0.5	0.2	0.0	0.1	0.0	0.1	0.0
uk	7.4	0.1	8.0	1.0	0.2	0.6	0.0	0.3	0.1
ca	4.3	0.1	1.2	2.6	0.1	0.2	0.0	0.2	0.0
in	0.8	0.0	0.3	0.1	0.6	0.1	0.0	0.1	0.0
de	1.1	0.0	0.5	0.2	0.1	0.3	0.0	0.1	0.0
cn	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
au1	1.0	0.0	0.5	0.1	0.0	0.1	0.0	0.3	0.0
br	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.2

Table 1: *The table represents a conversation matrix across geographies with columns representing posts and rows representing comments on them. Though conversations are biased locally, they do cut across English speaking areas. Language barriers appear to hinder interaction across certain areas.*

$A(u)$ is defined as:

$$A(u) = \sum_{v:v \to u} H(v)$$

And

$$H(u) = \sum_{v:u \to v} A(v)$$

The correlation between these ranking techniques is evaluated by the cardinality of the set intersection of nodes ranked to the same threshold. Figure 11 plots two independent correlations in increments of ten, upto a rank of one thousand. HITS Authority rank and PageRank is compared against the in-degree metric. The correlation between these rankings indicates that in-degree can be a good approximation of authority in a closed, controlled and generally spam free environment. *Authoritative, or socially important users are typically considered as thought leaders within the blogging community*.

A powerful hub is one that points to many powerful authorities. In the context of internal blogs, this could have important implications. A user who is a powerful hub, is also a good *social connector*, one who has followed and engaged in conversations and is aware of authoritative sources and hosted content. This motivates the identification of such connectors. Using the same approach we used for correlating authority scores, we plot HITS Hub score against out-degree in figure 12. Not surprisingly, the correlation seen in authority rankings extend to hub rankings as well.

4.4 Crossing Geographical Boundaries

Since blog adoption is global, the question of *Do blogs work as bridges across geographies* is of significance. To analyze this property, we augmented each node in the graph with the geographical location, and extracted edges where both the source and the destination are among the top nine contributing countries to the user base. Results are depicted in table 1. The row represents the destination node and the column represents the source, in other words the geographies of the post author, and comment author in a conversation. Each entry encodes the contribution, in percent, to the overall conversation graph. Though blogs have bridged geographies that speak a common language, conversations connecting Asia to the rest of the world remains limited, possibly hindered by language barriers.

4.5 Edge Betweenness Centrality

Betweenness centrality [13] measures the significance of nodes and edges as it relates to their centrality in information flow through the network. It hence forms an important measure for identifying effective word of mouth channels within a community. To identify if edges that reflect conversations across geographies are central to the

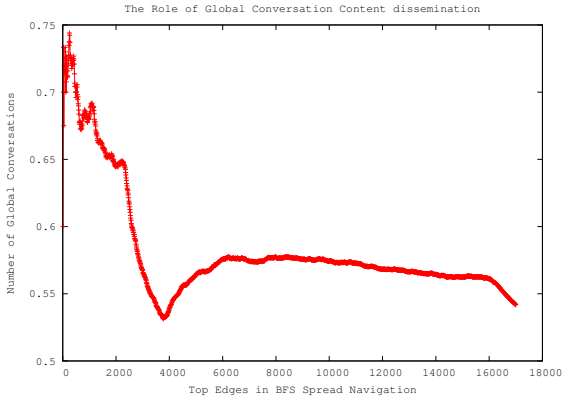

Fig. 13: *A high number of cross geography conversations among high ranked central edges shows the value of cross-geography interactions.*

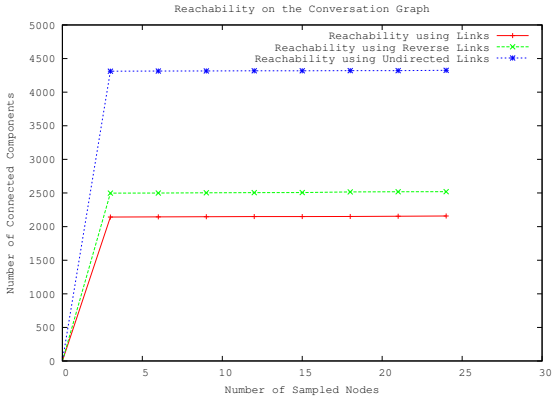

Fig. 14: *Reachability of the graph shows a highly connected component covering almost the entire graph. The disconnected components are constituted by certain team blogs that were active for a short time-frame.*

network, we rank edges based on their centrality, computed as:

$$C_B(u \to v) = \sum_{u \neq s \neq v \in V, u \neq t \neq v \in V} \frac{\sigma_{st}(u \to v)}{\sigma_{st}}$$

where $\sigma_{st}(u \to v)$ is the number of shortest geodesic paths from s to t that pass through the edge $u \to v$, and σ_{st} is the number of shortest geodesic paths from s to t.

Using a ranked list of central edges, we plotted the distribution of edges that cross geographical boundaries. As seen in figure 13, the high ratio of such cross geography edges among the top ranks show the value of global conversations. Such edges form significant bridges to information dissemination across a global organization.

4.6 Reachability

Reachability analysis is used to understand the structure of the network as it relates to its connected components. A strongly connected component on a directed graph G is a set of all nodes such that for any pair of nodes u, v there exists a path from u to v. The same applies to an undirected connected component on an undirected graph G^U. A well known implication of an analysis of connected components has been the identification of a bow-tie model on the Web Graph [7].

We identify connected components in the network through a BFS (Breadth First Search) traversal using a seed set of randomly sampled nodes. BFS is run on the graph G as is, on G^T obtained by reversing all edges, and on G^U obtained by making the graph undirected. As shown in figure 14, the graph consists of a strongly connected component of 2500 nodes that covers half of the graph, and an undirected connected component that covers almost the entire graph. Nodes which were not part of this giant undirected connected component, featured users who had posted a few times, and were lost by attrition without being sufficiently involved in the community. The properties of these disconnected components, offers both an opportunity and a challenge i.e. how can newly appearing disconnected components be encouraged to connect with the giant core component, improving blogger retention.

5. Modeling Enterprise Reach

We would like to capture the intuition that in a conversation, the relative position of employees part of the exchange, evaluated using a corporate hierarchy, can be significant to understand the reach and spread of ideas. We hence model an employee hierarchy as a rooted named unordered tree, from here on referred to as T. The root of the tree is the head of the organization. Each employee-manager relation is represented using a parent-child relation making managers internal nodes in the tree, and all non-managerial employees leaves.

We briefly introduce the readers to some basic tree properties. A node is an ancestor of another node u, if it exists in a path from u to the root node. The height of a node u in T, denoted as $h(u, T)$ is the distance between the node u to the root of the tree, with the height of root node being zero. The Lowest Common Ancestor (LCA) of any two nodes u and v in a tree is the lowest node in T that has both u and v as descendents. We define a sub-tree $T^{u,v}_{LCA}$, as a tree rooted at the LCA of u and v and featuring only nodes and edges that are in the path from u and v to the LCA. $E(T)$ is the set of all edges in the tree T.

The reach of a conversation between two users (employees) u and v is determined by the properties of $T^{u,v}_{LCA}$. We define one such property the reach, $R_c(u, v)$, as:

$$R_c(u, v) = |E(T^{u,v}_{LCA})|$$

$R_c(u, v)$ has a value of zero in a self-conversation, value one in a conversation between an employee and manager, and value two in a conversation between employees working for the same manager. Intuitively, reach is captured by the distance between two users in the corporate hierarchy, measured by the number of edges in $T^{u,v}_{LCA}$.

Using $R_c(u, v)$ as an atomic computation the normalized reach of a blog post R_p made by a user u, and hosting comments from a set of users V, can be computed as:

$$R_p(u, V) = \frac{\sum_{v \in V} R_c(u, v)}{|V|}$$

While reach captures the distance between two employees, it does not incorporate the aspect of spread when combining multiple conversations on a blog posts. To model this, we use *spread*, defined as the number of edges in the union of all conversations around a blog post. Spread is defined as:

$$S_p(u, V) = \frac{|\bigcup_{v \in V} E(T^{u,v}_{LCA})|}{|V|}$$

The distribution of normalized reach and spread across all blog posts is shown in figure 15. The reach of posts peaks around the

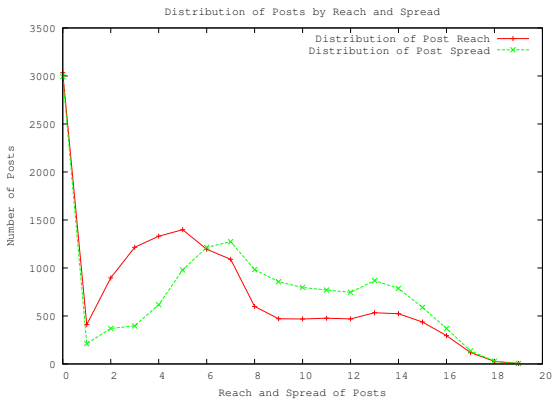

Fig. 15: *A balanced distribution of conversations based on their reach, show that conversations are not just limited locally between employees and their peers, but more spread out across the organization.*

value of four, and the spread peaks at around six. Both these distributions suggest that conversations are high across users working in close hierarchical proximity, and less exclusive among peers, and between employees and their managers. The persistent tail of this chart shows how blogs have enabled conversations across the organization.

These properties of conversations can be used to complement existing techniques, or to develop new techniques for identifying popular posts and tags, or to identify experts. In what follows, we briefly discuss a few such possibilities.

5.1 Ranking Posts

The overall rank of a blog post p can now be computed as a weighted sum of the number of comments (in-degree), its normalized reach and spread across the organization.

$$Rank(p) = w_c * |V| + w_r * R_p(u, V) + w_s * S_p(u, V)$$

where w_c, w_r and w_s are the weights attributed to the total comments, reach and spread respectively.

5.2 Popular Themes

The value of tags is as good as the posts they are associated to. For a tag t, we compute their aggregate value as a weighted sum of the occurrences and rank of posts they are attached to:

$$Popularity(t) = w_o * n(t) + w_{rs} * \sum_{i:t \in tags(i)} Rank(i)$$

where $n(t)$ is the number of times a tag is used in the application and w_o is the attributed weight. The second term is summation on the rank of all posts that are associated with t, and is weighted using w_{rs}

5.3 Finding Experts

The expertise of specific users on a topic t can be computed using aggregate rank of posts made by the user on topic t:

$$Expert(u, t) = \sum_{i:t \in tags(i), author(i)=u} Rank(i)$$

We are developing prototypes to evaluate the utility of these techniques for internal business intelligence. We will report results from our evaluation when they become available.

6. Related Work

The role of new knowledge management and communication tools is receiving widespread attention. Nardi et al [26] have studied the motivations for blogging in general and Wagner [30] and Grudin [14] have clarified the role of wikis and blogs within organizations and the usefulness of one over the other. Though we have shown how an organization's policy is critical to high adoption internally, McArthur et al [23] have explored how verifying for policy compliance can be automated. This could be important for public facing blogs.

Tagging and the concept of folksonomy is widely studied. Millen et al have discussed this in the context of an enterprise bookmarking service [24]. The use of tags in blogs has been studied by Brooks et al [8]. Farrell et al [12] have proposed tagging people and co-workers within an organization. Tags as a way to identify experts within an organization has been explored by John et al [17]. However existing work has not incorporated relationships between tagged entities (users) as proposed by us.

Complex networks have been analyzed in various contexts [27]. The social aspect of blogs has motivated recent research on analyzing networks materialized through blogs. Herring et al [16] have studied network characteristics of the general blogosphere. Adar et al [3] have explored information epidemics and ranking on the blogosphere, and Kumar et al [20] have looked at community dynamics and growth. Marlow [21] has studied the role of links between blogs using blog-rolls and permalinks as a metric to popularity. Adamic et al [2] have identified communities within the political blogosphere and analyzed conversations across these communities. Moor et al [10] have explored how conversations in blogs are different from those in other forms of communication. The role of materialized social networks as compared to employee hierarchy and their implications to the *small world* has been previously explored [1].

In our work we have focused on a broader study, to quantify multiple attributes of internal blogs that include structure and its use. We have also attempted to address questions on how characteristics of internal blogs and that of their hosted conversations can be useful for tools that extract business intelligence within an organization.

7. Conclusion and Future Work

Many corporations have internal blogs in use. As of this work, the structure and properties of these blogs were not empirically studied. While traditional approach of ranking entities (post, author, tag) still applies, our approach of utilizing the employee hierarchy to quantify reach is novel in the context of corporate internal blogs and should complement existing approaches well.

Research around social networks in general offers many interesting challenges [18], and our continuing research aims to address some of them. We are now focusing our study on the network characteristics of internal blogs, specifically on how the more explicit social models of employee hierarchy interplays with those materialized through blogs and how blogs are enabling a flatter organization.

8. Acknowledgements

We would like to thank users in the internal blogosphere for their timely comments and suggestions. Part of the analysis is based on conversations with them. We would also like to thank IBM's blogging team for sharing internal data for our analysis.

9. Trademarks

both. Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both. Other company, product, and service names may be trademarks or service marks of others.

References

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, New York, NY, USA, 2005. ACM Press.

[3] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit Structure and Dynamics of Blogspace. In *WWW 2004, Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[4] A. E. Alter. Emerging mobility, collaboration, and business process technologies, 2006. [Online; accessed 06-December-2006; `http://www.cioinsight.com/article2/0,1540,1957593,00.asp`].

[5] A.-L. Barabasi. Emergence of Scaling in Random Networks. *Handbook of Graphs and Networks*, pages 69–84, 2004.

[6] A.-L. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[7] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[8] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM Press.

[9] J. Cass, K. Munroe, and S. Turcotter. Corporate blogging: Is it worth the hype?, 2005. [Online; accessed 30-November-2006; `http://www.accountingweb.com/library/corp_blogging.pdf`].

[10] A. de Moor and L. Efimova. An Argumentation Analysis of Weblog Conversations. In *Proceedings of the 9th International Working Conference on the Language-Action Perpective on Communication Modeling, LAP 2004*, 2004.

[11] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66, 2002.

[12] S. Farrell and T. Lau. Fringe Contacts: People-Tagging for the Enterprise. In *WWW 2006, Collaborative Web Tagging Workshop*, 2006.

[13] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[14] J. Grudin. Enterprise knowledge management and emerging technologies. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, page 57.1, Washington, DC, USA, 2006. IEEE Computer Society.

[15] M. Guy and E. Tonkin. Folksonomie, tidying up tags? *D-Lib Magazine*, 62(1), 2006.

[16] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. Conversations in the blogosphere: An analysis "from the bottom up". In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, page 107.2, Washington, DC, USA, 2005. IEEE Computer Society.

[17] A. John and D. Seligmann. Collaborative Tagging and Expertise in the Enterprise. In *WWW 2006, Collaborative Web Tagging Workshop*, 2006.

[18] J. Kleinberg. Social networks, incentives, and search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–211, New York, NY, USA, 2006. ACM Press.

[19] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[20] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[21] C. Marlow. Audience, structure and authority in the weblog community. May 2004.

[22] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata, 2004. [Online; accessed 06-December-2006; `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`].

[23] R. McArthur, P. Bruza, and D. Song. Policy Conformance in the Corporate Blogspace. In *WWW 2005, Workshop on Policy Management for the Web*, 2005.

[24] D. Millen, J. Feinberg, and B. Kerr. Social bookmarking in the enterprise. *Queue*, 3(9):28–35, 2005.

[25] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom graphs: Findings and implications. In *ACM CIKM 2006*, 2006.

[26] B. A. Nardi, D. J. Schiano, and M. Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 222–231, New York, NY, USA, 2004. ACM Press.

[27] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[28] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[29] D. Sifry. State of the blogosphere, october 2006, 2006. [Online; accessed 03-December-2006; `http://www.sifry.com/alerts/archives/000443.html`].

[30] C. Wagner. Wiki: A Technology for Conversational Knowledge Management and Group Collaboration. *Communications of the Association of Information Systems*, 2005.