

No Blog is an Island — Analyzing Connections Across Information Networks

Smriti Bhagat

smbhagat@cs.rutgers.edu

Graham Cormode

graham@dimacs.rutgers.edu

S. Muthukrishnan

muthu@cs.rutgers.edu

Irina Rozenbaum

rozenbau@cs.rutgers.edu

Hongyi Xue

xuehy58@cs.rutgers.edu

Abstract

There are multiple information and social networks among individuals, from telephone to email, web, blog, instant message (IM) and chat networks. Prior work has studied each of these individual networks quite extensively, including telephone networks [7], postal network [20], web communities [17], and so on. Still, what is of great interest is how these networks collide and interact with each other. Each individual has presence in multiple networks and it will be of interest to understand how the references and presence in one affects that in the others. Previously such studies would have been difficult to do since each source of data was often “owned” by a specific entity. Blogs now provide a unique, public source of data that naturally provides visibility into multiple networks. For example, blog entries cite web pages, blogs have friends and communities, as well as blog profiles that have links to email, IM and other networks. In this paper, we use the blogs as a starting point to pull in data about these multiple networks and study how these multiple networks interact with each other. While much of the connections are *within* specific networks, there is still a wealth of information, structure and form to the connections *across* these networks as our analyses show.

1. Introduction

Today, there are multiple information and social networks among individuals, both emerging as well as well-established ones. For example, postal and telephone networks are long established. Emerging networks include the web, blogs, and instant chat networks (IM). Prior work has studied each of these individual networks quite extensively, and that has led to fundamental observations on the structure of the networks. For example, the six degrees of separation was a principle that was formulated in the context of the postal network [20]. Similarly, studies on telephone networks have revealed the communities of interest that exists among telephone users [7]. Among emerging networks, detailed analyses of web graph reveals the existence of clear structure [17].

The point of departure of our work is the premise that individuals exist simultaneously on many such networks, and the interaction between these networks is of great interest.

Example. Consider a typical young Internet user. She may update her MySpace profile, then post a blog entry on Xanga to point to the latest set of photos she has uploaded to Photobucket, and also to embed the most recent popular videos seen on YouTube. All the while she may be having multiple concurrent IM conversation with friends, and perhaps even checking her Hotmail account in case any

notifications have come in. It is clear that with this growth in the use of social networks and electronic communications, we should no longer think of a profile on a site or a blog service such as MySpace in isolation, but as part of a large system linking multiple information networks. It does not make sense to focus in on just a single information network. Instead, in studying social networks, we must be aware of the unique nature of each network, and view the larger interconnect of these networks. □

The central challenge in studying multiple information networks and the links between distinct information networks is the ability to get data from different networks. Different entities own data about different networks (e.g., postal and telephone, or telephone and web). It is impossible and sometimes illegal to join these sources for anti-competitive reasons in cases when one company owns the data from multiple networks (such as telephone and cable). Not surprisingly, there is not much published prior work on the connections between multiple information networks.

In this paper, we address this challenging problem. We approach it with blogs as our starting point. These are publicly available from multiple blog networks such as LiveJournal, Blogger, Xanga, etc. Some of the prior work has studied these networks individually and drawn conclusions about all the users of such services [14]. For us, the connection across these networks can be obtained by parsing these sources and is already a good starting point for our study. We can now go further and study the connections between these networks and the world of webpages. We are not aware of any detailed prior study of such connections. We can then look at the links from blog profiles to the network of instant message, email and VoIP networks, and use them to join the data sets.

Precisely, we present the following study:

- We gather a substantial data set of blogs from LiveJournal, Blogger, and Xanga, together with links amongst them, links to web pages from them, as well as various profile information associated with these blogs including age, location and interests when possible.
- We present a detailed study of the link structures:
 - blog—blog (e.g., LiveJournal—Blogger), where there is significant intra-site (within) and inter-site (across) linking activity, in Section 3;
 - blog—web (e.g., Blogger—webpages), where we observe significant variations in the quantity and nature of links between sites, in Section 4; and,
 - blog—messaging network (e.g., Xanga—AOL IM, Yahoo! email, etc), where we see a surprisingly large number of accounts on different systems that are linked by a common ID, in Section 5.

In each case, we present further study of the trends when projected onto age and location to see the commonalities and variations.

This type of cross-information network seems new. We are fortunate that blogs provide such cross-link references, in particular, to instant messaging networks. We observe that different blogging networks appear to have users with quite different habits, demographics, cultural backgrounds, and interests. We uncover many relationships: a large number of different “cultures” of the bloggers who use different blogging sites; collections of “symbiotic” websites which depend on their larger “hosts” for traffic and provide site-specific services in return; and evidence for a large number of users linked to multiple accounts across sites.

We observe that there is a current trend for some sites to try to monopolize the interest of its users, and provide all feasible services under a common brand: observe how MySpace offers social networking, blogging and movie and music hosting. Widespread adoption of a centralized service could reduce the need for cross-network studies. However, we conjecture that such efforts will have the same success as attempts in the 1990s to create “portal” webpages which cater to every perceived need of the user: due to the inherent nature of links, any entity in any of these information networks that is successful will link to and be linked from many other networks. Hence we will continue to need to better understand the complex inter-relations of these information networks.

2. Data collection

In order to study connections across multiple information networks, we collected large quantities of data, principally sourced from blog hosting sites (blog sites for short). We collected data from three popular blog sites — Blogger (<http://www.blogger.com>), LiveJournal (<http://www.livejournal.com>) and Xanga (<http://www.xanga.com>). These were chosen as three popular hosts, each maintaining tens of millions of user accounts (about 12M in LiveJournal and 40M in Xanga). Collection took place in Summer 2006. Our data is drawn from two main categories: user profiles and blog pages. Profiles contain various personal information provided by the user, which, in addition to the user name, may include date of birth, gender, location, occupation, interests etc. They also contain cues to find the user in other information networks, such as an email address, home page, or Instant Messaging (IM) identity. We parsed this data to extract all available details for each user, where provided. Our collection of blog pages contain the most recent (“front page” entries) for each crawled blog, and additionally some of the archived entries. The data was parsed and used to construct a graph, where each node corresponds to a blog user and a directed edge between two nodes corresponds to a blog entry of one of the users having a link to the other user’s blog (or entry therein). We collected blogs and profiles of 250K users from Blogger, 300K users from LiveJournal and 780K users from Xanga. All of this data was used in the analysis presented. Our dataset is available upon request.

The data from each blog site was collected in two passes. During the first pass, an initial “seed” set of blogs and profiles was crawled by randomly choosing blogs (for each site this necessitates a different approach, which we detail below). The second pass was to expand from the initial seed set by crawling blogs and profiles referenced by the observed blogs but not yet collected. Since the number of blogs hosted on each site is huge, we did not collect every blog, and our data includes blogs that are “known, but uncrawled”, i.e. which are known to be blogs hosted by a particular site as they were referenced by a blog already crawled, but whose profiles were not collected. The results for the number of edges in the resulting

	Blogger	LiveJournal	Xanga
Distinct Domains	331K	289K	74K
Blog-to-Webpage edges	997K	1089K	895K

Table 1: Counts of extracted web nodes and blog-to-webpage edges

graph are shown in Fig. 1(a), where we distinguish between links to crawled and uncrawled (unc.) blogs. Each site required different approaches in order to collect, parse, and clean the data, which we outline next:

Blogger. To obtain the initial set of blogs from Blogger, we crawled user profiles starting at the URL <http://www.blogger.com/profile/UserID>: each user has a unique integer identifier or *UserID* in the range up to about 50 million. However, not every user has a “public” profile, so not all such URLs lead to useful information. We incrementally probed values in the ranges 10000000 – 10010000, 15000000 – 15100000 and 16000000 – 16200000.

Each successful profile page retrieval was parsed to obtain the URL’s of (Blogger) blogs owned by the user, which was used to retrieve blog entries. One important issue is that Blogger allows a number of users to maintain one blog collectively. To fit this into our model of each blog having a single owner, when such collectively owned blog referenced another blog, we added an edge in our graph from each of the collective owners to the referenced blog’s owner. Since according to our analysis, only about 2% of blogs are owned by more than one person, this does not significantly affect the generated graph.

LiveJournal. There are few convenient ways to exhaustively crawl LiveJournal. To create the initial set of seed blogs, we used the “latest posts” feature to identify recently updated blogs (<http://www.livejournal.com/stats/latest.bml>). In addition to personal blogs, LiveJournal supports blogs owned by communities; these were omitted from our data collection. We normalized the data to avoid duplication: the same blog can be referenced by different URLs, for example:

```
http://www.livejournal.com/users/BlogName,
http://BlogName.livejournal.com,
http://livejournal.com/~BlogName;
```

similarly, underscores and dashes are (usually) interchangeable. Thus we converted all LiveJournal links to a canonical form. Individual LiveJournal posts can be “friends only” (only visible to users who are recorded as friends of the user). Since most blogs with hidden posts also have many posts that world-viewable, we did not take any special action over blogs containing such postings.

Xanga. To create the seed set for Xanga we took advantage of the concept of “metros”: each metro corresponds to a geographical region in which users locate themselves. A metro has anywhere from a single user to hundreds of thousands of users listed within it. Since each metro has an integer identifier, we could iterate through the metro ID’s, and parse user names. As with LiveJournal, there are issues of normalizing links to Xanga blogs into a canonical form. Xanga also allows users to “lock” their blogs, which makes the blog visible only to a selected group of people. Since this hides all content of the blog, such blogs were not crawled.

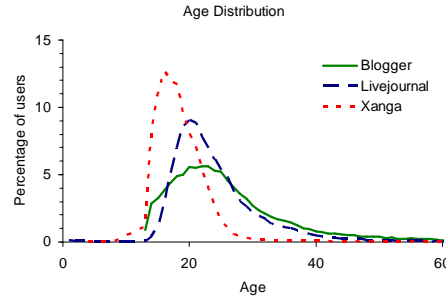
2.1 Properties of the data

From this data, we extracted information from profiles and blog entries, as detailed below.

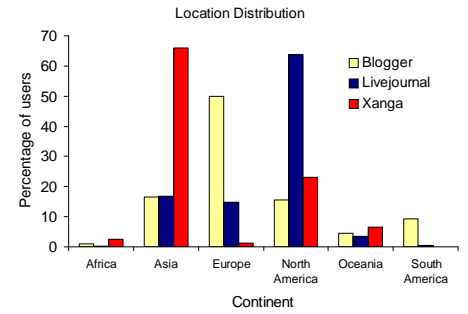
Friends. Both LiveJournal and Xanga allow users to add other users to their “Friends” category. In LiveJournal the number of friends and

	Blogger	LJ	Xanga
Blogger	190K	4K	236
Blogger unc.	234K	12K	1K
LJ blog page	1K	403K	316
LJ unc. blog	1K	227K	250
LJ friends		3,970K	
LJ unc. friends		7,489K	
Xanga	1K	1K	2,996K
Xanga unc.	29	55	1,065K
Xanga friends			88K
All	429K	12106K	4,063K

(a) Blog collection statistics



(b) User age distribution



(c) User location distribution (by continent)

Fig. 1: Data collection statistics and demographics

their blog names are listed in user profiles. We crawled the friends category of each profile in our data, which allowed us to add much more edges to our blog graph (see Table 1(a)). Friends in Xanga are more complex to extract: rather than presented as a webpage, they are displayed dynamically by javascript, requiring some site-specific code to extract a list of friends.

Weblinks. In addition to links to other blogs on the three servers, we also extracted links to other web pages that were referenced by the blogs in our data set. These included blogs on other servers and standard webpages. To simplify the analysis of those blog-to-webpage links, we only considered the domain names of the URLs. Table 1 shows the detailed statistics for the resulting number of distinct domains and the number of blog-to-webpage edges that were added to the graph as a result of this processing.

Demographics. Lastly, we looked at age, gender and location distribution of the users in our data set, in all three blog sites. Fig. 1(b) shows that Xanga has the youngest user population of the three networks, with the median age of 18 and the average age of 20 years old, while Blogger is represented by the most mature population among the three, with the median age of 24 and the average age of 27. Among the users that listed their gender, 63% of Xanga users are females while Blogger has only 47% of female users. LiveJournal does not make gender information available for public view. We grouped user locations extracted from all user profiles by continents, and the distribution of locations is shown in Figure 1(c). One interesting observation from the figure is that all three blog networks seem to have been adopted most on a different continent, Blogger having most of its users in Europe, LiveJournal in North America and Xanga in Asia. Naturally, we observed some amount of obviously false information in this demographic data: infeasibly high ages, fictional locations and so on. Nevertheless, from the observed distributions it seems that the vast majority of users are truthful when they give out information (we leave the task of quantifying this more precisely for future work). Our study is conducted to understand the features of these networks, and we do not reveal any personal information beyond the aggregate distributions.

3. Blog-Blog interactions

Bloggers interact with each other in various ways, for instance, they cite each other’s blogs, comment on postings, subscribe to blogs and befriend other bloggers. However, on closer inspection we discover that bloggers across different blogging services have very different communication patterns. We analyze the interactions of bloggers from the three blogging services that we study. These services are

popular among bloggers from different regions, age groups and interests. Even within these categories, we see quite distinct patterns in the interactions between bloggers across services.

3.1 Link statistics

One of the most visible form of interaction across information networks between bloggers is citations of other blogs. We first study “intra-blog” links: links *within* a single blog site. As shown in Figure 2(a), on average, a blogger cites between 10 to 30 (distinct) bloggers on the same server. For LiveJournal, we can break this down into around 25 links to “friends” (listed in the profile) and 5 links within the body of the blog (these may include links to friends as we did not filter duplicates between these two categories). Blogger presents no explicit mechanism for encoding friends, while Xanga automatically includes a list of the user’s “subscriptions” to other Xanga pages on the main blog page (which is similar to, but distinct from, Xanga’s notion of friends). In Xanga the average number of links to “friends” is very low - 0.1 links per user.

Observe that “friending” seems to have a large effect on the number of links seen in our data: the majority of links in LiveJournal are due to friends. Further, on investigating the links within Blogger, we discovered that the majority of these links are not within entries, but as more “static” content encoded in a blog-roll or link list as part of the user’s blog template. Thus, although there is no concept of “friends” within the Blogger service, some users seem to try to create the notion. The ease of adding friends, and other incentives (such as making it convenient to read the recent posts of those listed as friends) seems to explain the higher number of links within the LiveJournal and Xanga networks.

Popular blogs. We see more detail as we drill-down and investigate the distribution of the most cited blogs within each network. The distribution of blog citations is long tailed. For each server, less than 4% of all links are to the top 100 blogs (based on citation). Figure 2(b) shows the usernames of the ten most cited blogs in our dataset (for Blogger, we use the unique integer profile id). The striking differences in the nature of what is most popular on each blogging server gives a sense of the community of the users on each. The top blogs on Xanga from our data include blogs of celebrities, mostly from Hong Kong (MandyStarz, kellyjackie and stephy.tang). Bloggers that provide music codes to add to blogs which play music and video are also popular in Xanga (XaNgA_MuSiC, Music_Galore). Other “popular” users result from the popularity of such customization of pages: scripts that bloggers use to decorate their blogs include strings like ‘yourusername’ which are intended to be replaced with the user name of the blogger. Failing to do so creates a link to a

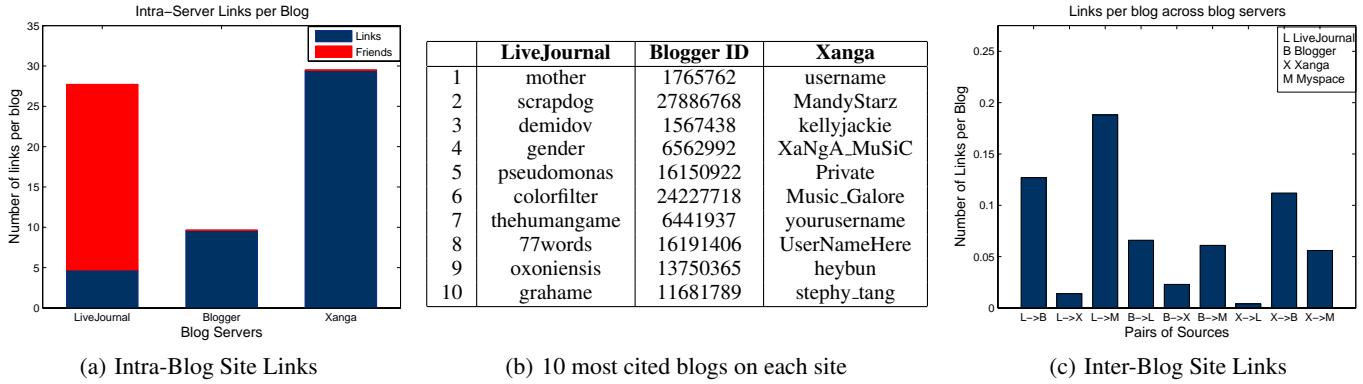


Fig. 2: Intra- and Inter- Blog site linking behavior

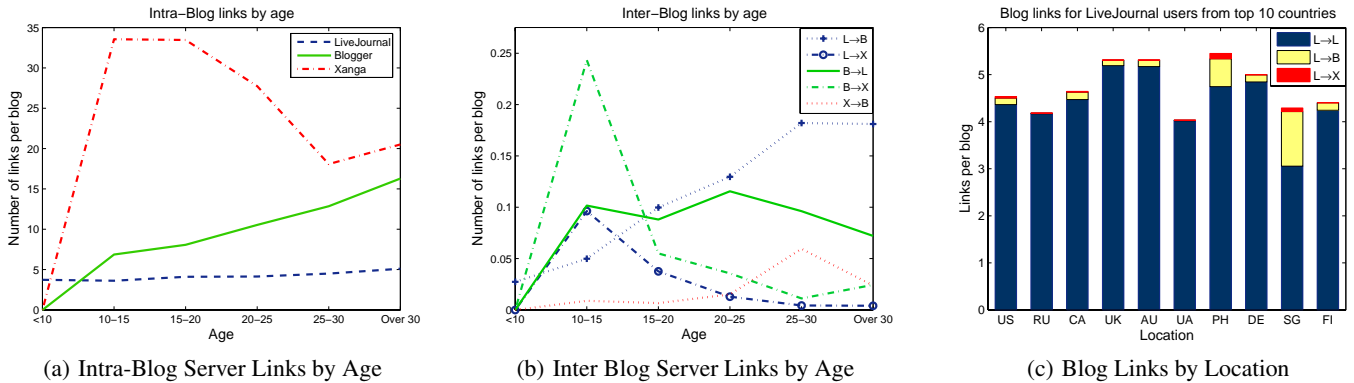


Fig. 3: Blog linking behavior broken down by age and location

blog which the owner has (either intentionally or by chance) titled ‘yourusername’.

The most popular blogs within LiveJournal also seem to reflect an interest in the site and beautifying one’s blog: the top ten list includes blogs of activists lobbying for changes (in the management and feature-set of LiveJournal), but is dominated by users who provide graphics, icons, and customized avatars. Blogger’s list of most cited bloggers was the most varied, and harder for us to classify. It includes users who keep a very large number of blogs on the system (in the hundreds), whose popularity may be explained solely by self-citations; in some cases, these are “spamblogs” (splogs), since the text of the blogs appears incoherent. Also highly ranked are members of a community of knitting enthusiasts, who make up a large and tight-knit group of Blogger users.

3.2 Inter-site links

We next study the patterns and distributions of links from one service to another: the *inter*-site links. Figure 2(c) shows the average number of links across blog sites. From our data, we observe that bloggers using the same blogging service cite each other significantly more than those using other services. Compared with the tens of intra-site blog links, the number of inter-site links are appreciably smaller, approximately 0.1 links per blog (totalling tens of thousands of links overall). Links to other social network and blogging sites are popular across all hosts—these could be links from a blogger to their own profile on MySpace, for example. Xanga and LiveJournal users show affinity towards Blogger users, who tend to make fewer blog citations overall. One notable feature is the minimal number of in-

teractions between LiveJournal and Xanga. This does not seem to be explained by age differences (although Xanga users are predominantly younger, LiveJournal users are not so much older). Instead, our working hypothesis is that this is more location/language based: Xanga users are predominantly from Asia, and often blog in Chinese, which makes interactions with LiveJournal users less likely, since they are predominantly from the US and Europe and write in European languages.

3.3 Projections on age and location

We further analyze the way users interact across information networks, by studying groupings such as location and age.

Projections on age. As illustrated in Figure 3(a) LiveJournal bloggers are consistent in the way they interact with other LiveJournal users across age groups. We cut off at age 30 since although we have ages that go up to 60s and beyond, blogs with age above 30 represent only a small fraction of users (less than 10% in all three datasets). We observe that teenagers on Xanga interact a lot with each other, while older users, who are fewer in number, seem to interact less; LiveJournal linkers are highly stable over age, while Blogger users seem to link to others more with age. Thus we see three distinct behaviors across different hosting sites, illustrating that it can be dangerous to make any sweeping statements about how bloggers as a whole behave based on observing a single site.

Figure 3(b) shows the inter-server interaction of bloggers. Teenagers on Blogger and LiveJournal interact with users of Xanga, which primarily has a teenage population. Similarly, LiveJournal and Xanga

bloggers in their late 20s and above interact with Blogger users, who are somewhat older on average. Links to LiveJournal users from those of Blogger were consistent across ages, and those from Xanga users were minimal.

Location. Shown in Figure 3(c) is the interaction patterns of bloggers from the ten countries with the largest population of LiveJournal users. Russian (RU) and Ukrainian (UA) bloggers prefer LiveJournal and tend not to interact with bloggers on other servers. There is more interaction between bloggers from Asian countries like Philippines (PH) and Singapore (SG) across servers. Bloggers from North America and Europe have moderate interactions across blogging sites. The tendency to link within site rather than across sites seems universal across all countries, influenced perhaps by the comparative ease of linking to another LiveJournal blog, and the factors which originally attracted the user to host their blog on LiveJournal (e.g. that many of their real life friends were already there). We omit breakdowns of other blog sites by location for brevity.

4. Blog-Web interactions

We next turn to how blogs interact with the rest of the web. For each blog, we extracted the links to different sites, and studied how these varied over different blog sites. Figure 4(a) shows the average number of web links found per-blog, and the average number to just the top 100 most common domains. We see that Blogger users link to the web much more frequently than Xanga users, according to our data; coincidentally, the average number of links to one of the top 100 most common domains (for that blog site) is about the same, approximately one per blog. Figure 4(b) shows the variation of linking behavior by age. We observe that the global habits seem to be consistent across ages for Xanga and LiveJournal, but for Blogger there is a marked increase in linking to the web as age increases. On the other hand, we saw no significant variation in linking patterns by location: we show the ten most popular countries in LiveJournal in Figure 4(c); other sites are similar.

4.1 Link categories

After categorization using existing lists such as the Open Directory project (<http://dmoz.org>) failed to yield meaningful or complete results, we performed a manual categorization of the top 100 domains for each blog site. We settled on ten categories for links:

- Links somewhat specific to blogs: hit counters and statistics (counters), personality quizzes (quiz), and icons, animations and “smileys” to be placed on the blog page (adornments)
- Hosting sites for media: images and videos
- Links to other blogs and social sites such as Friendster, MySpace (blogs).
- Other links: we broke these down into sites for news and sports (news), and those used to “reference” particular things, such as IMDB (to reference a movie), Wikipedia (a general reference) etc.
- What remained we divided into obviously planted commercial sites (spam) and the rest into the more generic class “content”.

The breakdown of the top 100 links (weighted by frequency) is shown in Figure 4. Across all sites, the links to other blog and social networking sites are about the same (8% – 12%), and links to media hosting of images and videos are also consistent (22% – 24% taken together). However, based on the nature of other links it is possible to characterize the “typical” user of each site as quite different. Within

Category	Blogger	LiveJournal	Xanga
Adornments	creative-commons	maps.google	gottem.net
Content	geocities	geocities	hk.geocities
Counters	statcounter	statcounter	xtracker.us
News	BBC news	BBC news	hk.news.yahoo
Photos	picasa	photobucket	photobucket
Quiz	bunnyherolabs	quizilla	bunnyherolabs
Reference	wikipedia	wikipedia	wikipedia
Social	myspace	myspace	myspace
Video	youtube	youtube	youtube

Table 2: Most popular link in each category across blog sites

Blogger, a large fraction of links are to news and reference sites, suggesting a keenness to point to and comment on the wider world. In LiveJournal, there is about half the fraction of links to news, and a large proportion of “quiz” type links, to a seemingly endless variety of personality quizzes. In Xanga, it is almost as if the outside world does not exist: news, references and links to other content make an almost insignificant fraction of the top 100 links (which in turn are about half the total links). Instead, almost a third of the links are devoted to graphics, animations, and other tools to “beautify” their page. Including the tools used to count and categorize their visitors takes the total to over half. Blogger users also frequently track their visitors, while in LiveJournal counters are rare.

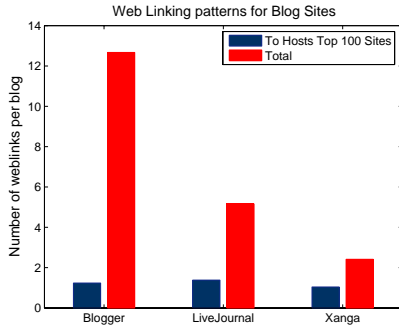
Popular websites. Further analyzing popular websites sheds deeper insight into the different blog sites. We pick out the most popular link in each of our assigned categories in Table 2. Some sites are universally popular: Wikipedia, YouTube and MySpace unsurprisingly dominate the reference, video and social categories. We see the bias towards users from Asia in Xanga by the popularity of Hong Kong news and hosting sites. Picasa is popular in Blogger due to the tool’s ability to make posting images to a blog easier. The Adornments category gives an intriguing insight: Xanga users adorn their pages with graphics and toys from a Xanga specific site called Gottem (a “symbiotic” site that depends on Xanga for its existence); LiveJournal users use a tool which embeds a map that shows the location of the blogger and nearby bloggers; and Blogger users include a creative-commons licence to regulate the sharing of content they create. This poses a natural question: given these characterizations of a “typical” user of each site, is it the case that a user is attracted to a site based on their personality, or that the way they use a site is shaped by the features that each site and its symbiotic sites provide?

5. Blog-IM interactions

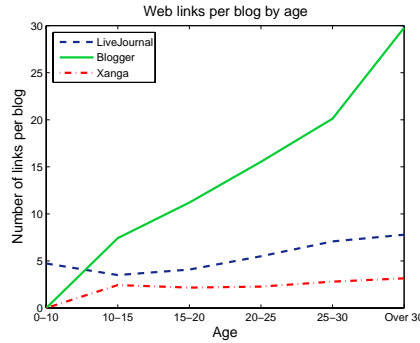
In addition to the web and other blogs, blog users typically interact on other electronic networks, such as Instant Messenger (IM) and email. Within the user profile there is typically room for the blogger to list several IM identities, an email address and a personal webpage. All three sites studied allow users to enter IDs for AOL Instant Messenger (AIM), ICQ, MSN, and Yahoo!, as well as an email and web address. Blogger additionally includes Google Talk, while Xanga offers Jabber; LiveJournal includes both of these and also IM/VoIP network Skype. In this section we study the prevalence with which this information is available, and use this information to understand the extent to which one user may create multiple blogs.

5.1 General profile statistics

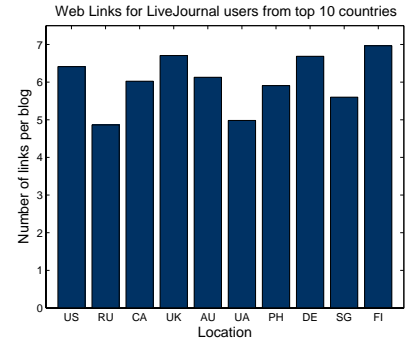
Figure 5(a) shows the percentage of the profiles for which users filled in the value of the category being examined from each of three blog networks studied. One can see that LiveJournal users tend to disclose a lot more personal information in comparison to other net-



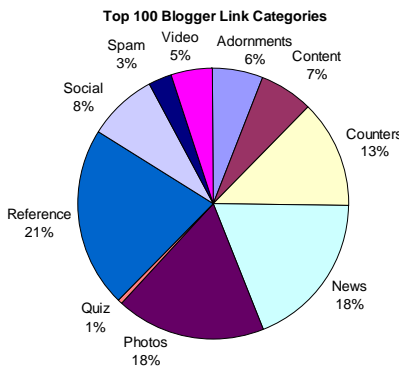
(a) Global Web Linking Frequency



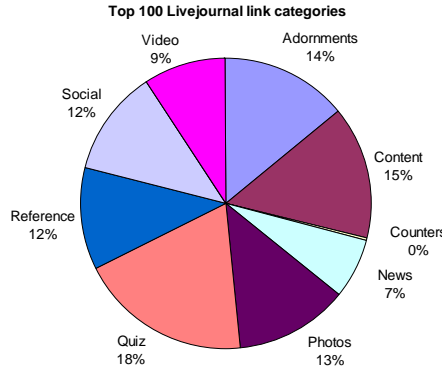
(b) Web Linking with Age



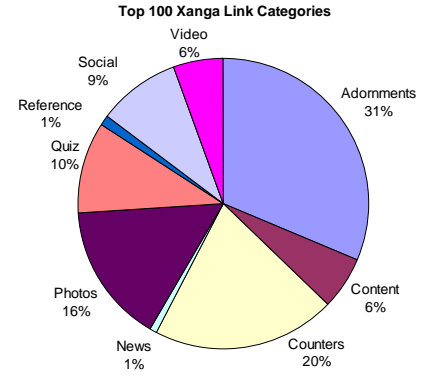
(c) Web Linking with Location



(d) Blogger web-link categories



(e) LiveJournal web-link categories



(f) Xanga web-link categories

Fig. 4: Web Linking Patterns and Categorized links of different blog sites

works, while Blogger users are the most discrete among the three networks: none of the examined Blogger users had listed and made visible their email address under the *Email* category. Over half of Xanga users list some URL under the *Webpage* category; however on closer examination the URLs listed we saw that a large number do not refer to personal webpages but rather to popular or favorite websites, (e.g. www.google.com, www.dictionary.com). Note that here we kept and compared the complete URLs. All categories suffer from a small percentage of “non-legal” textual responses, such as “ask me”.

Xanga treats email addresses differently: users can provide their email address to Xanga, and visitors can use the website to send email, without the address being visible directly. Thus, although over a sixth of Xanga users have provided email addresses, we cannot use it when trying to match users across networks.

Figure 5(f) illustrates that the percentage of users that share any IM contact decreases with age. This matches the perception that teenagers are IM savvy (and are happier to disclose such information). Meanwhile, the percentage of users who prefer email increases with age. More Xanga users share an IM contact than provide an email address (even though this email is not visible to others), whereas LiveJournal users consistently give a public email address more often than an IM.

5.2 Identity matches within a blog network

A natural question given the ever growing number of blogs is to ask how many represent distinct people, and how many belong to the same user. Given the anonymity provided by the Internet, it is hard to answer this with certainty, but we attempt to shed light on the

matter by studying the amount of duplication within our data sets.¹

To identify such blogs within the same blog network we compared values of each of the categories specified in a profile to the same category in all of the analyzed profiles within the blog network. If the value in two profiles is identical we declare a *profile match*. In many cases users list their instant messenger ID’s and email addresses under various categories regardless of the title of the category. For instance, the same email address can be listed under MSN and Yahoo! addresses. To record such cases we compared the value of each of the categories from a profile to the values of all other categories in different profiles. The additional matches found this way are shown under the category *cross-boundary*.

Figure 5(b) shows the percentage of matches from LiveJournal and Xanga relative to the overall number of profiles examined. The number for Blogger were negligibly small and so are omitted. Figure 5(c) shows the results for all three networks, relative to the number of profiles that have values listed for the category being examined. Both figures show a large number of matches on webpage in LiveJournal. This is due in part to a large number of people listing a LiveJournal community webpage under their Webpage category. There is also an appreciable fraction of non-personal websites as mentioned above.

In addition to the number of identity matches, we examined the number of profiles we considered to represent the same user due to a

¹ Note that the method of data collection may introduce some bias. The fact that we follow links from an initial set means that we are likely to find multiple blogs with the same owner that are linked together. This may lead to overestimating the quantity of duplication.

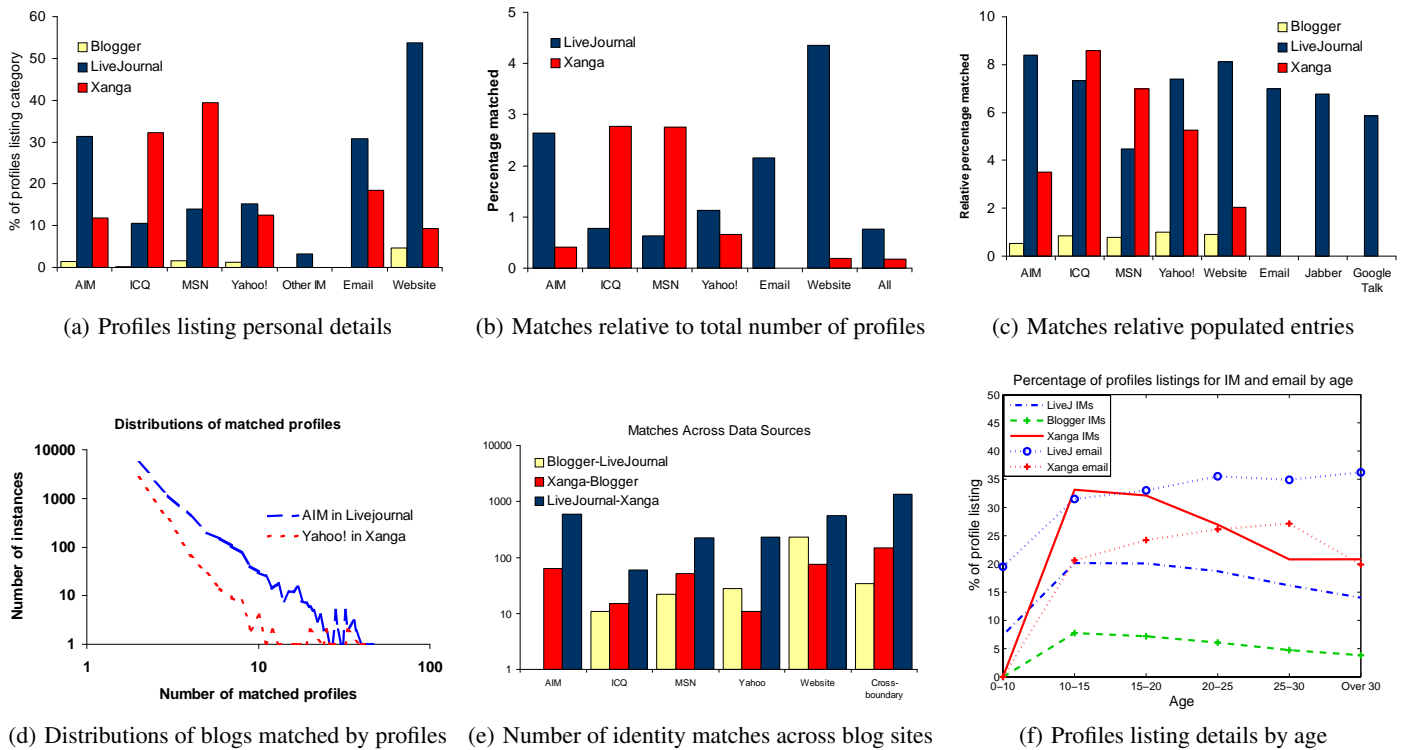


Fig. 5: Blog-IM Interaction Patterns

match in each of the populated profile categories. The distribution of the results looks very similar for all categories, thus we only present graphs for two categories in Figure 5(d). We see that the bulk of the matches correspond to 2 or 3 matched profiles, and the distribution resembles a straight line when plotted on a log-log scale, implying a power-law distribution.

The cases where a match was detected across a very large number of blogs are less likely to represent the same person, since they typically correspond to matches of the most commonly used invalid responses, such as ‘ask’, ‘ask me’, ‘n/a’, ‘yes’, ‘no’, etc.

Lastly we calculated the overall number of users that seemed to own more than one account within the same network according to our criteria. Our experiments show that in our data less than 1% of users from Blogger have multiple profiles, while for LiveJournal and Xanga this number is higher, about 4 – 5%.

5.3 Identity matches across blog networks

Next, we considered profile category value matches across different blog networks, using a method similar to the one described in the previous section. We examined only the profile categories common to all three networks. The results of this experiment are shown in Figure 5(e).

Not shown is the number of matches with a profile in all three blog networks: we found 2 valid matches based on Yahoo! IM and 9 cross-boundary matches. In general, the number of matches between pairs of blog networks were quite small: the most identity matches were found across LiveJournal and Xanga due to the large set of profiles analyzed and the amount of profile information filled in by these network’s users. These yielded a few hundreds of matches in each category; in contrast, we found a few thousands of matches in each category *within* each network. It seems that Xanga and LiveJournal

users often develop a certain amount of loyalty, so when starting new accounts for whatever reason, they stay with the same host. Meanwhile, Blogger has tens of matches both within and across networks, but note that Blogger also makes it easy to create multiple blogs under a single profile, giving an explanation for fewer profile matches under Blogger. Further investigation is needed to test these hypotheses, since our data, although large, is not exhaustive, and the further difficulty of judging user intent from the data.

In summary, our experiments show a surprising willingness of users to make their private contact information available. This is most common on Xanga which has the youngest users. We also see a noticeably high number of potentially duplicated profiles across sites, sometimes due to setting up multiple blogs (one for family, one for friends), perhaps due to wanting to “start over” afresh.

6. Related work

Over recent years, there has been significant analysis of weblogs, social networks, and electronic communications. We survey some of the most relevant works for our focus on blogs and their connections to other information networks. A general consensus is that there have been observable shifts in the way such blogs are perceived and used in the few years since user-friendly tools for creating blogs have been widely available [12], hence conclusions about network properties may change over time.

Kumar *et al* [14] give a crisp overview of how different age groups within LiveJournal cluster into different interests, and how communities of bloggers based on links grow over time [15, 14]. Two recent studies [21, 5] analyze features of blogs which correlate with the age of the blogger, such as punctuation features, number of friends and interests, and vocabulary choice (interestingly, they omit cross-network features such as IM address or weblinking behavior). Her-

ring *et al* [12] perform a manual analysis of two hundred blogs to refute the notion that blogs are principally “link lists”, and demonstrate that outside the top-100 “A-list” blogs, randomly chosen blogs are more likely to be “personal journals”. They categorize the about 90 links, and show that about half are to general websites, while 16 are to news sites and 13 to other blogs. Hurst [13] contrasted the difference between bloggers using Blogger and MSN spaces, and hypothesized reasons for the differences based on features such as the location of the bloggers. Most recently, a survey by the Pew Charitable Trust obtained opinions of randomly chosen US bloggers [18]. However, the nature of the sampling (random telephone sampling) meant the sample was small (233), and in particular included no bloggers under the age of 18—an age range which includes about half of the declared ages of the users in our data.

Where larger collections of data have been made, analysis has often focused on the graph properties of the induced graph of blogs with edges formed from web links or from explicit “friend” links. Kumar *et al* [16] showed that the friendship graphs from Flickr and Yahoo! 360 consist of singletons, small communities and a giant component, and gave models to generate such structures; [11] analyzed properties such as average path length and centrality of 5000 blogs found by crawling from four seed points. Other studies have looked at small collections of blogs from specific well-defined communities such as politics [1] and Knowledge Management [8], with some commentary on how these blogs interact with the larger web.

Generally, the question of how such networks relate to other information networks has been limited to how blogs link to the rest of the web. Typically web links have been treated as a short hand for content or ‘memes’ to be used as indicators for how information travels within the blogosphere, for example in [3, 2, 6, 9], without much consideration of what sites the links are to. The question of how links between blogs arise has also attracted significant study. Properties of the friendship graph from the Wallop system are used to analyze what makes users stay active [19], and what factors influence users to join in the first place [10]. Machine learning techniques have been used to identify which factors influence the formation of friendships in LiveJournal (and academic collaborations) [4].

7. Conclusions

While a lot of prior analyses of information networks have focused on individual ones such as postal, telephone, web or particular blog networks, we focused on studying the connections across multiple information networks. Starting from blogs provides a way to study connections across multiple information networks including blogs, the web, and IM. We noticed very high variation in the statistical trends of these connections across different networks, and when projected along age and geographic locations. We also discover significant links across information networks via email or IM identities. We believe that a more extensive study will help tease out and model the variations better, which will be the basis for future study.

Acknowledgments: This research was supported by a KDD supplement to NSF ITR 0220280.

References

- [1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *International Workshop on Link Discovery (LinkKDD)*, pages 36–43, 2005.
- [2] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, 2005.
- [3] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.
- [4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 44–54, 2006.
- [5] J. D. Burger and J. C. Henderson. Barely legal writers: An exploration of features for predicting blogger age. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- [6] E. Cohen and B. Krishnamurthy. A short walk in the blogistan. *Computer Networks*, 50(5):615–630, 2006.
- [7] C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. *Intelligent Data Analysis*, 6(3):211–219, 2002.
- [8] L. Efimova, S. Hendrick, and A. Anjewierden. Finding ‘the life between buildings’: An approach for defining a weblog community. In *Internet Research 6.0: Internet Generations*, October 2005.
- [9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6(2):43–52, 2004.
- [10] L. Gu, P. Johns, T. M. Lento, and M. A. Smith. How do blog gardens grow? Language community correlates with network diffusion and adoption of blogging systems. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- [11] 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 *Hawaii International Conference on System Sciences (HICSS)*, 2005.
- [12] S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *Hawaii International Conference on System Sciences (HICSS)*, 2004.
- [13] M. Hurst. 24 hours in the blogosphere. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- [14] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, 2004.
- [15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *International World Wide Web Conference*, 8(2):159–178, 2005.
- [16] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 611–617, 2006.
- [17] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. The Web as a graph. In *Principles of Database Systems (PODS)*, pages 1–10, 2000.
- [18] A. Lenhart and S. Fox. A portrait of the internet’s new storytellers. Technical report, Pew Internet & American Life Project, 2006.
- [19] T. Lento, H. T. Welsch, and L. Gu. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *3rd Annual Workshop on the Weblogging Ecosystem*, 2006.
- [20] S. Milgram. The small-world problem. *Psychology Today*, 1:61–67, 1967.
- [21] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.