

The LJSmallWorld Project: Extracting Social Networks from LiveJournal

Kathryn Hagan <kmh2124@columbia.edu>

April 27, 2007

ABSTRACT

LiveJournal is a very popular blogging / social networking website where connections to others are both explicit ("friend" / "friend of" lists) and implicit (mentions of other users in entries, comments given/ received). This paper describes a program, LJSmallWorld, which aims to examine a corpus of LiveJournal data and graph the social networks involved, locating cliques and hubs, and drawing conclusions about the strength and possibly even nature of the relationships.

MOTIVATION & RELATED WORK

Web logs, often called "blogs", are Internet journals that have gained popularity explosively in recent years. The extraction of social networks from these journals is currently a very active area of research; it has been suggested that analysis of such data would be useful for content-based advertising, gaging consumer reactions to products or political candidates, or even understanding more about the dynamics of social groups in general (both online and off).

Most notable for this project is the work of Mishne and Glance, who gathered and analyzed blog comments and conclude that they are an often-overlooked source of valuable data regarding social networks, as well as being a fairly reliable indication of the popularity of the author¹. Adamic and Adar extracted social networks from user's individual homepages, stating, among other things, that their work suffered from a lack of explicit connection data; they say that "a future direction for this work would go beyond homepages to obtain social links directly from users."² In a separate paper, Adamic and Glance point out the growing importance of blog analysis in their study of discussion topics related to the 2004 U.S. Election³.

Though many bloggers publish their journals individually, a number of collective blogging sites have become popular for their combination of blogging with social networking. One of the largest of these is LiveJournal⁴, a site which, at the time of publication, has over 1.8 million active accounts⁵. In addition to its popularity, LiveJournal is especially interesting for this project because it provides two levels of connectivity between users. One, which can be considered an explicit connection, is each user's "friend list" – a list of users whose entries they can read from their LiveJournal page, and who can be given access to non-public journal entries. The second level, the implicit relationship between users, is the actual interactions that take place – mentions of other users in entries or informal discussions which take place in the comment area of each entry. Comments also serve as some indication of which journals a user is actually reading.

IMPLEMENTATION DETAILS

LJSmallWorld uses a corpus of data retrieved from the LiveJournal website as needed and cached locally. For fetching and parsing data from the site, it uses the LWP::Agent and XML::FeedPP⁶ modules. For construction and analysis of the network, it uses the Clair library⁷, specifically the Clair::Network module. The two modules that make up the LJSmallWorld back end are LJ.pm, which wraps LiveJournal requests, and LJNetwork, which is a representation of the network itself. Graphs were generated using the Pajek graph analysis tool⁸.

When asked to generate a network, the program will first request a random username from LiveJournal if one was not provided. Then, it uses the `get_users_in_graph` function to generate a list of the nodes which will appear in the final graph, downloading friend data from LiveJournal as it goes. The users are then added to the network, with the connections between them (the explicit connections from their friend lists) set to a low initial value. (Note: Though the algorithm does more work than necessary, as it adds connections and nodes outside of the graph, this implementation was chosen so that subsequent requests from within the same network would be faster.)

Next, LJSmallWorld examines the recent entries of the users in the graph in order to determine the relative strength of the connections. It does this by scanning each entry for instances of the string "lj:user=`username`", which is the format of an internal LiveJournal link – in other words, a link to another user's journal. These links can be placed within the entry text by the journal's author, as well as being embedded in the page automatically when another user leaves a comment on that entry. Each reference to another user in the entry text increases the strength of the edge from the author to that user, while each comment from another user increases the strength of the edge in the opposite direction (from the user to the author).

Finally, a subnetwork is generated, containing only the requested users and edges between them. This subnetwork can then be analyzed or written to file.

SYSTEM EVALUATION & RESULTS

In order to verify the accuracy of the graphs and tune the default weights, graphs were generated for a set of users whose relationships were known (the web surrounding the researcher's journal). After correctness had been verified, graphs and statistics were generated for a group of random users.

Figure 1 shows one of the smaller random graphs, with the edge weights and vertices labeled. Darker vertices imply stronger connections; very light ones imply weak connections. Figure 2 is the same user as Figure 1, but at the next level out; this user is somewhat unusual in that her friends are not very strongly interconnected, each having their own separate group.

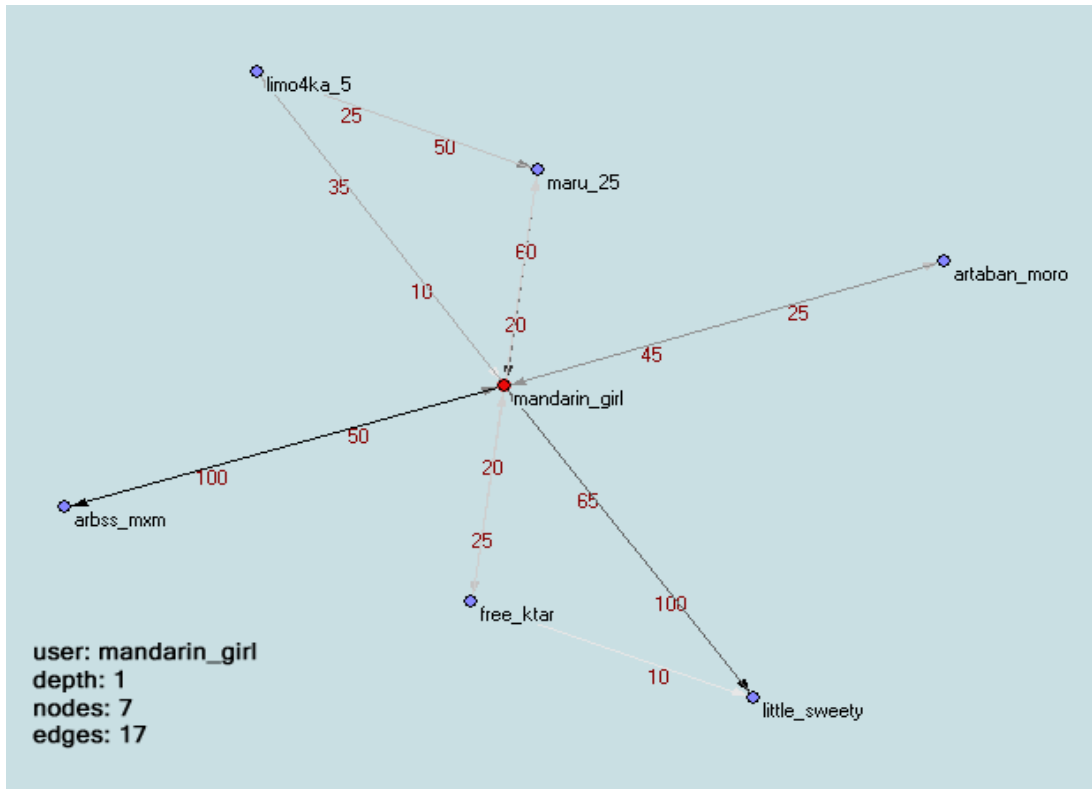


Figure 1. mandarin_girl's network at level 1

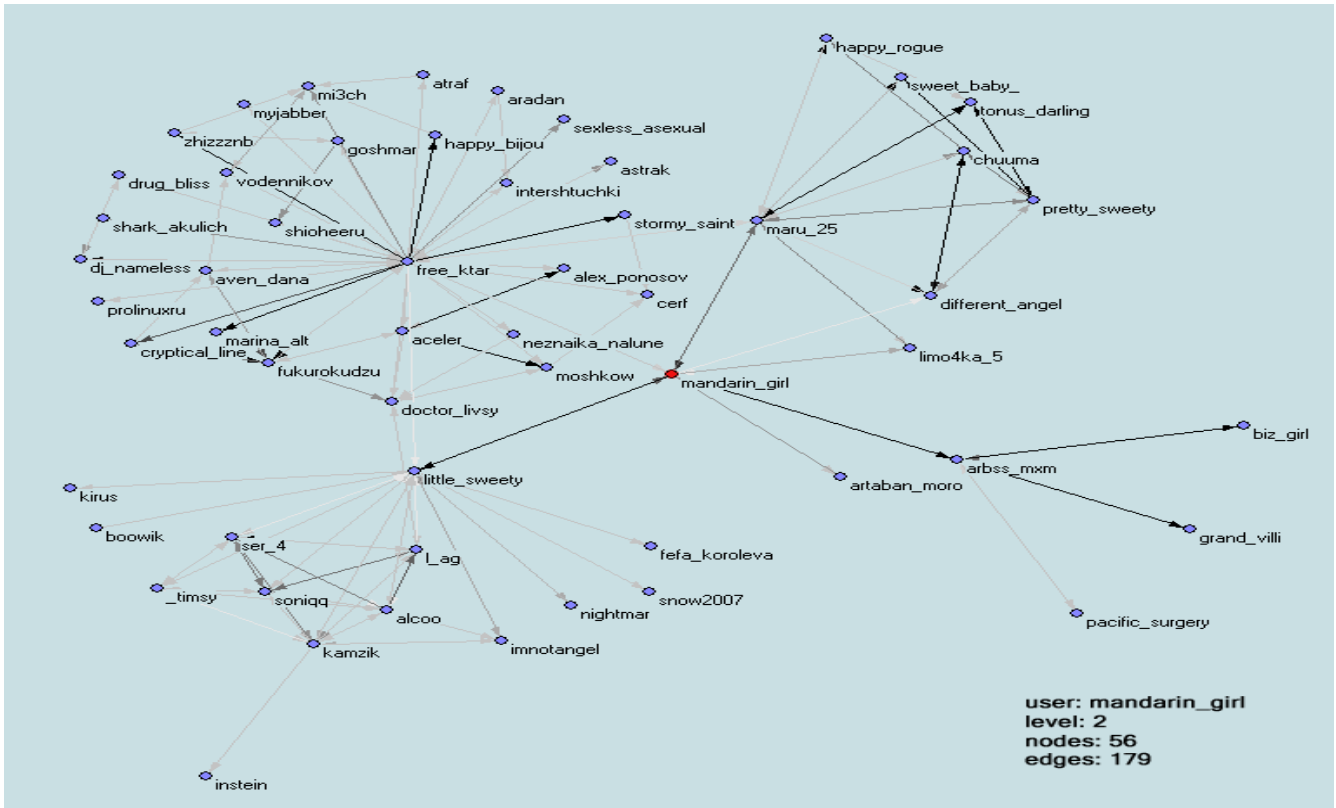
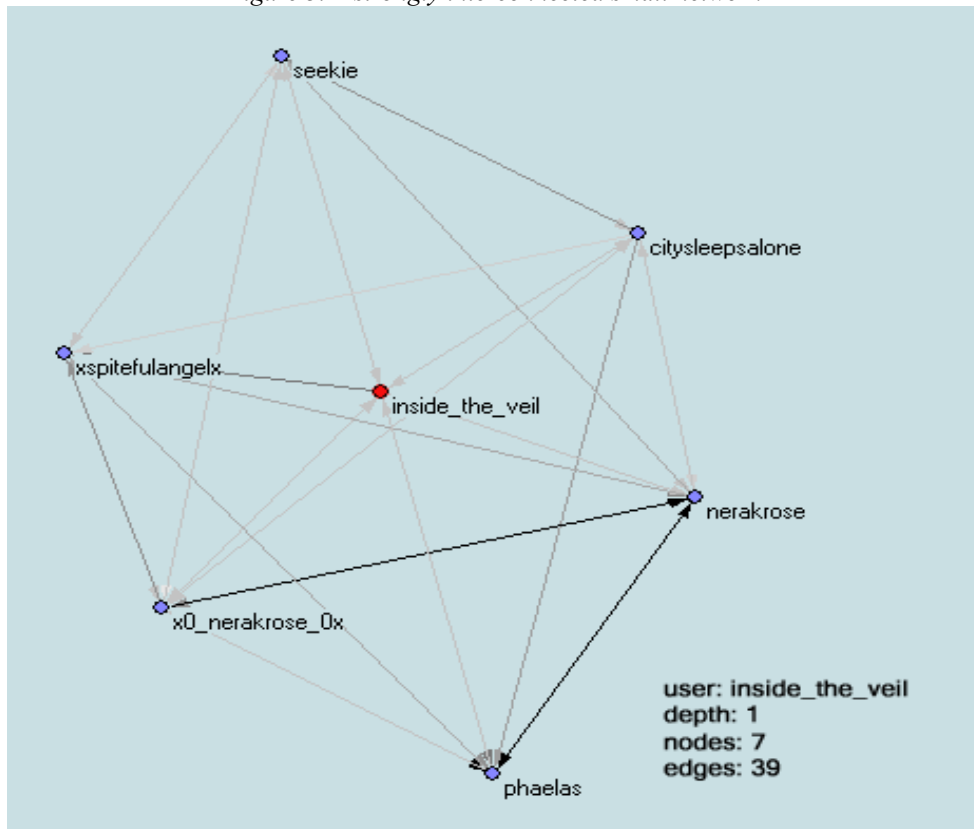


Figure 2. mandarin_girl's network at level 2

In contrast, some of the graphs were very strongly interconnected, such as the small one in Figure 3.

Figure 3. A strongly interconnected small network



The final three graphs demonstrate the variety that occurred among the larger networks.

Figure 4. A large network with a strongly-connected center and a fringe of outliers

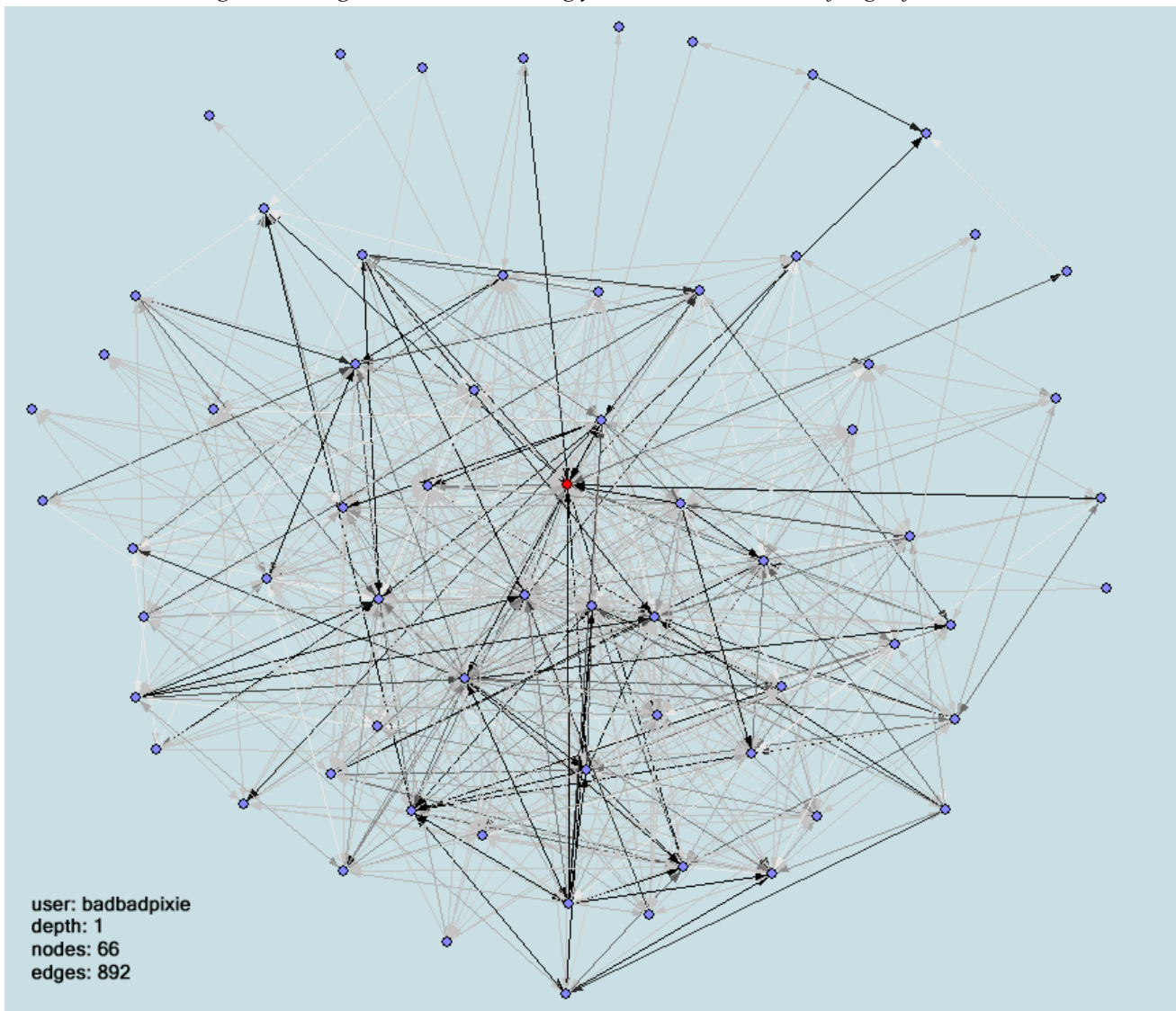
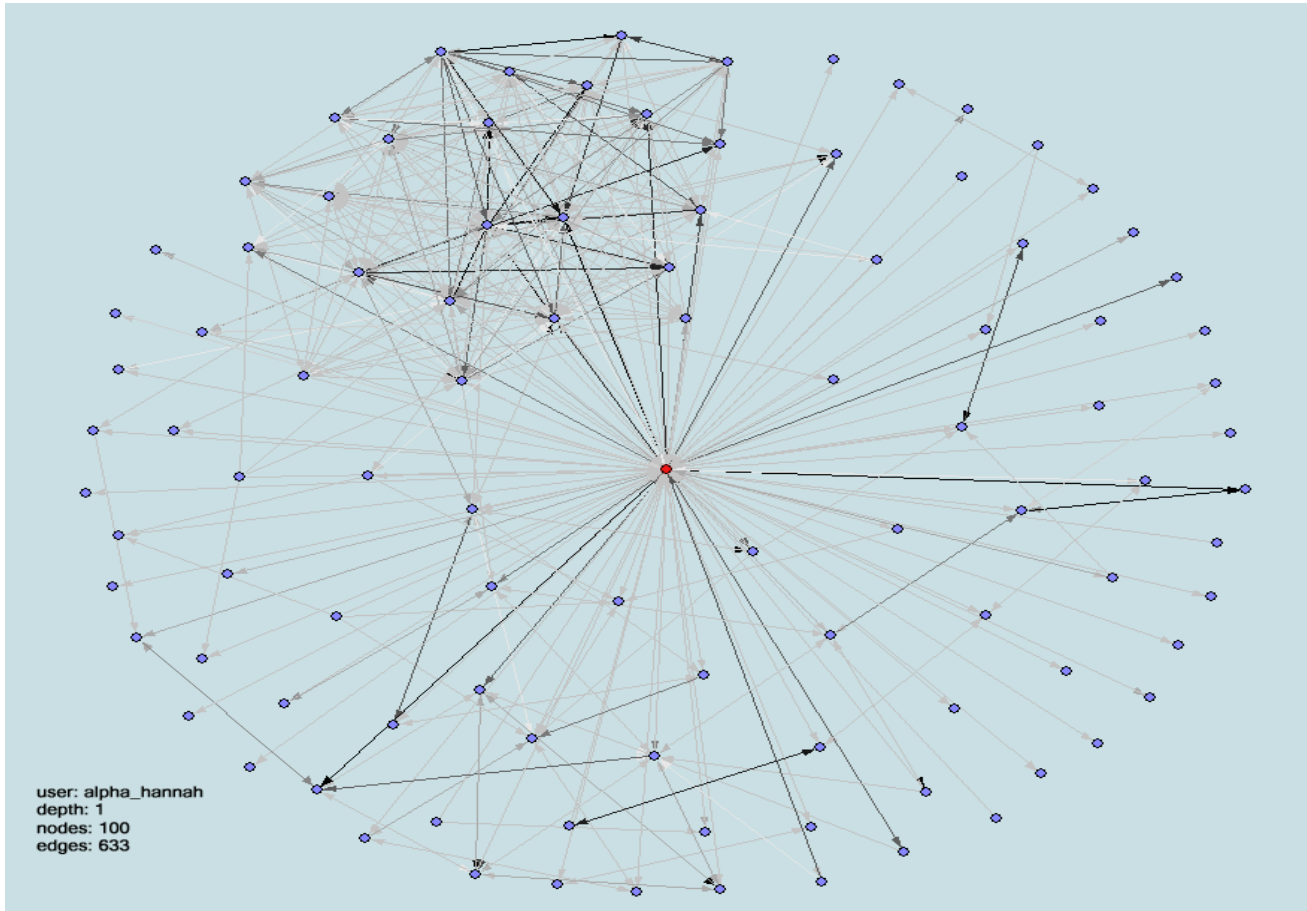


Figure 6. A very sparsely-connected network with a strongly-connected cluster



Finally, the following table sums up the basic characteristics of these networks.

Table 1. Network Statistics

Graph	Nodes	Edges	Diameter	Avg. degree
mandarin_girl 1	7	17	2	4.86
mandarin_girl 2	56	179	6	6.39
inside_the_veil	7	39	2	11.14
badbadpixie	66	938	4	28.42
alectoerinyes	99	325	4	6.57
alpha_hannah	100	633	4	12.66

The two major problems with the system are speed and scalability. Unfortunately, LiveJournal requests that automated requests be limited to five per second. This is the major source of the speed problem when the data is not yet cached in the corpus, as LJ.pm has a built-in sleep for 0.2s after every request to the server. (A secondary source of slowdown is the Clair::Network function for generating a subnetwork from a full graph, which can sometimes take quite a long time to finish, especially on large networks.)

The system also fails to scale very well to large amounts of data or large networks. The local cache of friend data and entries to disk is slow and the files become untenably large very quickly. One way to ameliorate this problem might be by using a database back end to store and retrieve cached data rapidly. With more storage space, generated networks and subnetworks might also be stored to disk.

A more minor difficulty is that many entries are not public, and many users do not make public entries at all. Though this could be solved for the web surrounding a specific user by implementing digest authentication, I chose not to implement it due to the privacy concerns that would arise. Also, the system does not recognize “threading” in comments – in other words, that users may actually be responding to other commenters instead of to the original poster. A system which compensated for these distinctions might be able to differentiate user relationships at a finer degree of granularity, but would probably have no real advantage, in the large scale, over the current system.

FUTURE WORK

In addition to entries and comments on those entries, more information, such as geographical location or interests in common, is available which could be analyzed in order to determine which users are closest. Heuristic analysis of comment and entry text could be used to determine the topics most often discussed amongst groups of users, or the nature of the relationships between users. At a more user-oriented level, a more efficient version of this program, equipped with a database backend and authentication abilities, could be converted to a web-based application that would allow individual LiveJournal users to generate and view their own graphs.

REFERENCES

1. G. Mishne and N. Glance. Leave a Reply: An Analysis of Weblog Comments. In WWE (Workshop on Weblogging Ecosystem), 2006.
2. L. A. Adamic and E. Adar, 'Friends and neighbors on the Web'. *Social Networks*, 25(3):211-. 230, 2003.
3. Adamic, L. A. and N. Glance, "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." WWW2005 Conference's 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics, 2005.
4. LiveJournal, <http://www.livejournal.com/>.
5. LiveJournal Statistics, <http://www.livejournal.com/stats.bml>.
6. Available on CPAN at <http://search.cpan.org/~kawasaki/XML-FeedPP-0.21/lib/XML/FeedPP.pm>.
7. The Clair Library, <http://tangra.si.umich.edu/clair/clairlib/>.
8. Pajek: Program for Large Network Analysis; available at <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>.