Understanding Pure Social Networks: Structure, Connectivity, and Patterns of Interests*

Chris Tanner, Chu-Cheng Hsieh, Keenahn Jung†
Department of Computer Science
Los Angeles, CA 90095
{christanner, chucheng, keenahn}@ucla.edu

ABSTRACT

In this paper, we attempt to better understand social networks, as they currently represent an alarming majority of all web site traffic. Specifically, we aim to measure the structure, connectivity, and patterns of interests that may emerge. We feel that better understanding such may not only lead to improving social networking sites, but could also lead to developing new personalized social applications and mediums, improving targeted marketing, and assisting in sociological research. Unlike most related research efforts, we concentrate on a *pure* social network–Facebook.com–in the sense that our medium of choice is one whose entire premise is based on developing social networks. This contrasts with many other *loose* social networks such as photo or video sharing sites (i.e. Flickr, PicasaWeb, YouTube, Google Video) which are primarily focused on providing featured content, but allow for social-network-like communities to form as a complemented feature.

We developed a comprehensive *crawler* for Facebook. We attempt to accurately model the global structure of Facebook by focusing our crawl on the regional network of Los Angeles, California. We confirm that the vast majority of users are part of a large strongly connected component (SCC), whereby any user can reach any other user via 4.8 friendship hops on average. Moreover, we demonstrate that this is a lower-bound for regional networks, and that global connectivity on Facebook should be comparable. In addition, we show that the distribution of friendships adheres to the power-law property. We devise an efficient algorithm to find commonality of interests between all users, and show that friends are almost twice as likely as non-friends to share common interests with one another. Last, as an added bonus to the research community, we will make our extensive crawling data publicly available to academia after we remove sensitive, personal data and obtaining web storage.

Keywords

social networks, web crawling, Facebook, connectivity

1. INTRODUCTION

Social Networking web sites consume a large percentage of today's internet traffic, and are amongst the most frequented web sites in the world: As of October 19, 2007, Myspace.com ranks at #6 and Facebook.com at #7 [1][5]. However, little research concerning this global phenomenon has been conducted. As the World Wide Web began to blossom exponentially, it sparked the attention for much research. Specifically, attempts were made to graph the structure of the web [12]. This research work paved the foundation for future work involving crawling algorithms, searching, data extraction, community discovery, and much more. Ultimately, it was this initial investigating exploration that helped raise attention and harness revolutionary efforts.

Consequently, it seems natural for efforts to be put forth toward understanding the ubiquitous social networking phenomenon. Specifically, we believe that this understanding could possibly lead to great advances, akin to those that were facilitated by the aforementioned web structure research. Realistically, it is at least feasible to garner useful information that could lead to the development of new social applications and mediums, and could lead to improving targeted marketing. For example, if we could model users' interests and find insightful correlations, it would be useful to develop social systems that are personalized towards its users. In fact, during our research, Facebook introduced a new ad system that allows users to broadcast their purchased items to their friends [7].

Moreover, we believe that accurate models of social networking could provide much towards the field of sociology. For example, the typical, fundamental question that one thinks of when concerning sociology is how many people separate you from a particular celebrity (or random person) in terms of friendships. This six-degrees of separation topic has undergone much research since its infancy in 1967[21][17]. A common difficulty for such experimental research is often obtaining a wealth of accurate, objective information. For

*This research is still in progress and more results will be added later

†Chris' tasks included research focus, crawling trials, and all writing; Chu-Cheng developed code for all experiments, which he also conducted, and was responsible for general research and management; Keenahn did the enormous task of writing the highly complex crawler–the crux to making this all possible

this given question of degrees of separations, it would be infeasible to sample millions of individuals and ask them to provide a comprehensive listing of all of their friends. However, with the new prevalence of social networks, this wealth of information is present and can aid in avenues that extend beyond the separation question. For example, Yahoo! and Google have each worked on returning query results based on social networks to which the user belongs [10][3]. In addition, there have been recent efforts to use social networks to aid in improving search engine results [18] and filtering e-mail spam [15].

Specifically, our goal is to model the global system of social networking web sites, while obtaining statistics related to structure of friendships, connectivity strength, and patterns of interest. Naturally, to obtain such information, we must initially crawl users' pages. Reports have shown that Myspace.com has anywhere from 100 million to 200 million active registered users[4][6], with Facebook.com in second place with at roughly 80 million. Naturally, we must focus on a subset of users, and later we explain why we chose Facebook.com's regional network of Los Angeles, California (L.A). Note, though, that assert that this provides a lower-bound guarantee towards connectivity.

To the best of our knowledge, we are one of the few to research a *pure* social network web site–one whose entire premise is to construct a social network. (Many other *loose* "social networks" are merely featured-content based sites who provide a platform from which user communities may form.) Moreover, most pure social networks (i.e. Myspace and Facebook) provide users with privacy controls. Sadly, it seems that many users actually use them, consequently allowing only their accepted friends or fellow community members to see their information. We circumvented this obstacle by creating accounts and joining the Los Angeles regional network, thus gaining permission to view the mass of semi-private accounts within Los Angeles. We believe we are the first to take this approach.

Our contributions are:

- Our highly customizable, efficient, robust crawler (it can even handle network connection resets).
- Our crawled dataset–the majority of all Facebook.com users in Los Angeles, California
- Confirmation of the small-world and power-law distribution characteristics, when concerning friendships
- Statistics concerning users' interests, including the fact that friends are almost twice as likely to share interests than strangers are.

The rest of our paper is as follows: In Section 2, we provide a general background of social networks and their terminology. In Section 3, we detail our development of our crawler, explaining our motivations, limitations, and methods. We mention our efforts toward understanding patterns of users' interests and the underlying structure of friendships in Sections 4 and 5, respectively. Our latest experiments and findings are presented in Section 6. The myriad of potential research for future is listed in Section 7, and we summarize in Section 8.

2. BACKGROUND

If you have ever used a social networking site, you can probably skip this section. Yet, to be self-contained, we provide the following overview:

Although social networking sites may differ greatly from one another, as seen by the aforementioned contrast between *loose and* pure social networks, much of their underlying concept remains constant: each active user represents a unique entity that represents oneself. Each user has a customizable *profile*, whereby the user may provide a picture of himself and list characteristics to represent who he is in the real world. Social networking sites typically provide a general template of these characteristics, allowing users to enter their favorite activities, interests, music, movies, and general information about themselves. Optionally, users may choose to leave any of these fields empty. In addition, users may add their own fields to their profile. In fact, recently, there has been an enormous burst in the development of "applications" for users' profiles. These applications are essentially displays of enriched content, which often make use of and/or allow input from others. Along these lines, profiles are also becoming increasingly rich in content, for users may typically upload pictures, videos, and other media that includes them or interests them.

Now that a user has a profile, he may want to set privacy restrictions so that not everyone can view his information. For example, if the user lists his home address, phone number, or embarrassing/incriminating pictures, he obviously may want to protect such from the public eye. Social networks allow for communities-like sub-networks to form. These networks are generally founded to represent members of a shared university, geographical region, company for employment, or interest. With this, users may conveniently set their privacy controls based on each of the networks to which they belong.

Now that we have discussed a user's profile and the existence of sub-networks, we lead to the key element of social networking: friends. Users may often search or browse for other users. Upon finding other users, generally one would want to view the found user(s) profiles. This may or may not be possible, depending on the aforementioned privacy controls. However, a user would not want to accidentally limit a long-lost friend from viewing his profile, preventing them from re-discovering each other again. So, usually users have somewhat relaxed privacy controls–specific enough not to allow *everyone*, but general enough to allow people who belong to the same networks. We mention this because it plays a key role in devising ways and knowing what is possible for us to crawl. Nevertheless, users may find each other, and *friendships* may form: user A making a request to be friends with user B. User B may choose to accept or decline the friendship request. Note that friendship is always bi-directional: if user A is listed as one of user B's friends, then user B will be listed as one of user A's friends.

With our mention of users and friendship, it should be clear that we can represent a social network via a set of graphs:

- Let vertices V_1 and V_2 represent two users. V_1 and V_2 are directly connected via edge E if they are friends with each other
- Construct the connections $\forall\ V_i \in V$, where V = all users and V_i represents a unique user

3. CRAWLER

3.1 Limitations

Ideally, we would wish to crawl the entire network of Facebook's 80 million users. As mentioned, due to privacy limitations (some people only allow their friends to view their profiles), computation resource limitations (80 millions user profiles would consume roughly 8 Terrabytes of data), and crawling prohibition (Facebook bans accounts that crawl), this was not possible. So, how do we choose to sample? Also, in addition to Facebook's efforts to detect crawling, occasional network issues may arise: how we do handle the case when our local internet briefly looses connection? What about when Facebook is busy and simply cannot respond in time? What do we do when we encounter users who hide their listing of friends? We address these questions below.

3.2 Approach

Random sampling is a possible approach. However, unless you are certain that you can crawl a significant portion of a given network, the results will likely skewed and disconnected. For example, you may randomly pick user A who has 100 friends, then randomly pick user B who has 100 friends, none of which is A. Then, we randomly picked user C who has 100 friends, neither of whom is A nor B. Clearly, with this approach we cannot accurately model the structure of the network unless we have high coverage. Consequently, we chose to base our crawl on a Breadth-First Search (BFS) approach: we will merely be crawling friends of friends of friends, etc. This snowball effect presents a bias towards the connectivity issue [19][16], for all users we explore are clearly friends of someone, so the average degree of separation is misrepresented as being lower than the actual.

In order to combat this, we restrict our crawling to users belonging to the Los Angeles, California regional network on Facebook. We remove all links to users who are outside of our chosen sub-network. So, if a user within our chosen subnetwork has 100 friends, but only 10 of which are part of our network, we have greatly limited sequentially-following possibilities. Although this does not completely eliminate the breadth-first search bias, it can significantly limit it, for it is not uncommon for people to have 0 friends within a given regional network to which they belong. We can justify this by stating: (1) for an fast-paced, large city, it is common for people to quickly move to and from it, so most of their friends could exist in other cities; (2) a user's listed Facebook friends is almost always a lower-bound for their number of actual friends–a socialite may simply choose not to frequently use Facebook; (3) 9.1% of our crawled users had privacy settings that prevented us from viewing their list of friends. Not only do these realities help limit our bias, but keep in mind that we are focusing on the entire city of Los Angeles. Los Angeles's population of 3.8 million places it as being the 2^{nd} most populated city in the United States[8]. So, if we could crawl a majority of it, the bias would appear to be almost none-existent; regardless of how we crawl, if we cover a significant portion of it, we are accurately represented a large environment where two randomly chosen people will doubtfully know each other.

We chose Los Angeles partially due to its aforementioned property of being large. More importantly though is its Facebook's lack of popularity within Los Angeles: Facebook started at Harvard University in Cambridge, MA. Naturally, due to its social networking basis, the popularity of Facebook predominately grew within the New England States. Still, the majority of Facebook users and web traffic stem from the Eastern Coast of the United States[2], as seen in Figure 1. As we can see, per capita, Los Angeles has a very weak presence on Facebook. So, we can assert with confidence that the connectivity results we obtain are a lower-bound for that of more typical Facebook networks.

Our approach was essentially based on BFS, but extended to suit our needs. We start with a random list of ID's in the Los Angeles network. We obtained these ID's from the "Browsing friends and people from the Los Angeles Network" page with random ordering. Let us call this set of ID's U. Formally:

Let U = $u_1, u_2, \cdots, u_n$ Let list V keep track of visited ID's (initially $V = \emptyset$) Let list Q represent our current BFS queue.

Our algorithm follows:

```
While U != {}:
 Remove random element u_i from U and add to Q
 While Q != {}:
   Dequeue element q_1 from Q
   Save profile page for q_1
   Add q_1 to V
   Foreach friend f of q_1:
     Add 1 to friend count of q_1
     If f is member of Los Angeles network:
       Add 1 to Los Angeles friend count of q_1
       If f not in V and not in Q:
         Add f to end of Q
         If f in U:
           Remove f from U
```

Since network connections are unreliable, we put in a number of fail-safes. Whenever we try to open a page using our python browser, if we encounter an HTTPError we retry up to 10 times with a delay of 5 seconds between. If we ever get logged out of Facebook, we detect this and attempt to login again. If a more serious error occurs, we wait a minute and then restart the main loop again. Also, we periodically serialize the state of the program and write it out to disk. The state of our crawl is simply defined by the elements in our various lists. We would write the state to disk after each 25 profiles crawled. Note: these integer values concerning time and serializing are all customizable.

In later iterations of our program, it became necessary to split up the Q and U lists so that we could distribute it to multiple machines and have them crawl in parallel. When we did this, we lost the ordering of the BFS queue, but this is okay since we ended up crawling everything anyway. When

Figure 1: Facebook users per capita, translated downward.

we split up the Q and U lists, say into $Q_1, Q_2, \ldots, Q_m$ and $U_1, U_2, \ldots U_m$ for m splits, we also introduced a new list for each split called N. N contained Q - Q_m. In other words, it contained all the ID's in other BFS queues.

4. CONTENT/INTERESTS

4.1 Motivation

As mentioned in the Background Section, user profiles are becoming increasingly rich in content, often including social applications that allow for interactions between a user's friends. Many of these applications seem to be of an entertainment nature, but they are still often somewhat personal. For example, Facebook users commonly have profiles that contain hundreds of pictures (and videos) they have taken. Users can tag these files, display their blogs, and essentially provide everything about them in front of public display. We suggest that obtaining and making use of this information would allow one to have the largest database of volunteered, *personalized* data in existence.

Of this wealth of information, which should we obtain and what should we do with it? Facebook's general template has a field labeled *Interests*. Since all users are directly prompted to optionally complete this data, we figured this would provide us with a large, useful data set. With this data, we thought it would be interesting to:

1. Determine what are the most popular interests for Los Angeles. We could then compare this to other regional networks and see what regional characteristics form. For example, from a marketing perspective, it would be great to glean that people in Los Angeles are mostly interested in Photography, Movies, and Running, while people in Boston are most interested in Education, Red Sox, Music.

2. Measure how common is one's interests on average. For example, if we pick any interest from any user, on average how many other people share that interest?

3. Determine if this commonality of interests is significantly higher among pairs of friends vs amongst everyone. One would suspect so, but if this is not the case, it might make for an interesting sociological research topic.

4.2 Limitations

Of course, not all users provide their interests, and even if they did, privacy settings may prevent us from viewing their interests. This availability of data was not a *huge* problem, but it could potentially skew our ability to extract meaningful patterns. The bigger issue, however, is the natural language processing problem: we are ultimately interested in the *meaning* behind the words. If different users list "running" and "runnin," it is hard for us to realize that these are the same interest. Plus, this is a syntactical issue; more interesting situations are when different users may list: "jogging through the streets," "jogging," "jogging!," and "jogging, but only on a treadmill." Even as a human it can be difficult to determine how specific we should be when classifying. Last, even if we only look for keywords, there could be errors: "school," "anything but school" are clearly not the same. Related to this is that some users have different definitions of "interests": instead of succinctly listing each interest separated by commas, some users occasionally are more free with their writing: "i love to hang out with friends and enjoy myself. i love to read when i feel like it. i love watching movies, and talking on the phone, and going swimming. the one thing i really like is to just hang out with friends. feel free to talk to me and get to know me better. peace!"

4.3 Approach

We decided not to tackle this battle of correcting users' input and distinguishing between semantically equivalent but syntactically different words, for this problem is an entire field in itself. We believe our choice will not significantly skew our results, for there is enough users and commonly shared, well-formatted interests for us to work with. So, we merely extract all interests from all pages by delimiting each interest item based on punctuation. We then remove superfluous punctuation and convert all words to lowercase for the sake of being case-insensitive. We look through all users' interests and remove those interests that only have one occurrence, for we feel we can safely deem these as being bogus listings. Our assumption is that there are enough people within Los Angeles that nobody has a unique interest. Even if this were the case, the consequence of wrongly deleted a few unique elements is much less severe than including the vast collection of bogus, ill-formed interest items.

To answer (1) from above, we simply iterate through each user's interest. For each encountered interest, we add it to a HashMap, where the key is the *interest* and the value is the *# of occurrences thus far*. This gives us our needed data with O(n) time-complexity, where n = total # of interests among all users.

To answer (2), we provide an efficient O(n) solution that only requires two passes through the entire database of interests (see 4.3). Basically, we first store a count of how many people like each of the interests, as we did in (1). Through this pass, we also store each user's interests. Now, we merely need to go through each user, looking at each of his interests, and check how many other people share that interest–which we can lookup in constant time via our hashmap of interests-count.

To answer (3), the solution is similar to (2), but we must only look at interests among all the friends of a given user.

Repeating this for many (or all) users provides us with a measurement of how much friends have in common, on average.

```
hashmap userInterests (user -> list of interests)
hashmap interestsCount (interest -> #)

// store interests in memory
for each users
  userInterests.put(user, list of interests)
  for each interest the user has
    interestsCount.put(interest, currentval +1)

// eleminate bogus singletons
remove all singleton interests from interestsCount
remove all users who now have no interests


// count # of shared interests
totalShared = 0
for each user
  sharedIntr = 0
  for each interest the user has
    sharedIntr += interestsCount.get(interest) -1
  totalShared += sharedIntr/(# of user's interests)
totalShared /= (# of users)(# of users -1)
print totalShared
```

5. STRUCTURE/CONNECTIVITY

We aim to model the global structure of Facebook by focusing on the subset regional network of Los Angeles. We are motivated by the typical sociology question of "how many degrees of separation are there between two random strangers?"[17] We remind the reader that one's Facebook connections are a lower-bound to the actual number of friends one has in real life; thus, whatever our findings are, we can confidently state that the actual degrees of separation are much lower in real life. With our crawled data, would be interesting to define a metric of *influence* or *importance*. With this, it would be ideal from a marketing perspective to target these more influential individuals. A simple approach would be to merely define *influence* as the aggregate sum of values based on his friends, where each of these values is also an aggregate sum of their friends. This basic equation seems akin to the PageRank algorithm for search engines[11]. For now, we simply focused on determining: (1) the average degrees of separation between randomly chosen Facebook users; (2) the distribution of friendship; (3) general statistics such as radius and diameter of the large strongest connected component.

To answer (1), it was pretty computationally expensive, for we needed to efficiently store the friendships of each user. Nevertheless, we devised a solution and were able to accurately measure our golden question. As for (2), it has been found that power-law distributions (see below equation) are often the underlying behavior of many models, including the link structure of the web and the social network Orkut [19][12]. We were curious if our results would yield the same. Initially, this seemed doubtful, for although we mentioned that many users may have few friends, it would still seem reasonable to suspect that the majority of users have a handful of friends. Thus, our distribution of friendships would take on a bell-curved shape.

$$p(x) \propto \frac{1}{i^x} x > 1 \quad (1)$$

6. EXPERIMENTS/RESULTS

6.1 Crawling

Remember that our crawler not only downloads each user's profile, but it obtains a listing of each user's friends. This requires many more page requests, as Facebook only displays 10 friends per page. As a consequence, we can only download approximately 10 user profiles per minute (which involves roughly 1 page request per second). When we started crawling the Los Angeles network, it had roughly 320,000 users. So, we would need 22 days of flawless crawling. Soon, we realized that Facebook activitely monitors accounts that crawl, for it soon became a cat-and-mouse game. (being the mouse is not fun). We eventually created 26 alternative accounts for crawling, 18 of which were killed. After utilizing various lab machines and having friends (in various states) generously run our crawler, we were able to successfully crawl **301,692 profiles.** In fact, we reached the end of our original queue. So, it appears we exhausted the entire SCC that encompasses the vast majority of users. Note that these 301,692 profiles were not all unique, and that despite our efforts to minimize duplicate crawling, we were left with **176,786 unique profiles.**

By the time we ended our crawling, the Los Angeles network had grown beyond 350,000 users. This 10% growth further demonstrates our argument that many users Los Angeles users are new to Facebook and may naturally have few listed friends. With 350,000 users, it appears we only had roughly 55% of user's pages, yet we somehow reached the end of our crawl. How is this possible? The answer is that: (1) recall that many users have few friends. With only a few friends, it is easy to have a tiny self-referencing group of 3 or 4 users; (2) this is especially exacerbated by the fact that we ignore non-Los-Angeles friendships–increasing the misleading # of few friendships; (3) many users hide their friends list, even from fellow Los Angeles users. In fact, we found that 9.1% of our crawled users did such.

In fact, our result that 55% of users belong to the SCC is surprisingly large: [19] found that removing just 1% of the highest-degree nodes from a SCC yielded them with a SCC that was roughly 60% of its original size. Recall, we could not reach 9% of our users' information due to privacy settings. If only 11% of these privacy-restricted users are of very high-degree, then we can easily see that **at best our SCC will be 60% of its realistic size.** Our results show that even with limited connections, we still survive a large SCC, which agrees with others' findings [13].

6.2 Interests

There are a myriad of experiments to conduct on user's profile content. Yet, due to time constraints, we stuck to our mentioned 3:

1. Most common interests within Los Angeles: See Table 1. The only surprising result is #4, dancing. These

Table 1: Most popular interests within Los Angeles

#	interests
1	music
2	movies
3	traveling
4	dancing
5	reading
6	shopping
7	sports
8	art
9	basketball
10	friends

terms are naturally generic, common interests, but we view worthwhile to obtain this listing. Now, it will be interesting to see how other cities compare to this.

2. The likelihood that someone else will be interested in any randomly chosen user's interests. In other words, if we pick any interest from any user, what is the likelihood that another randomly chosen person will share that interest? **Our results show that the likelihood is .015** So, regardless of your interest, roughly at least one person out of every 100 will also share your interest. Of course, this is again a lower-bound, for we remind the reader of our strictness in matching only identical interest listings, and that many users either list no interests, or they set privacy restrictions

3. The likelihood that friends share common interests: this is similar to above, except we only consider direct friends to be candidates for matching interests. **Our preliminary results show that the likelihood is .023.** So, roughly 2 out of every 100 of your friends will also share your particular interest.

   Also, we state that average person listed 8.26 interests.

6.3 Structure/Connectivity

Within our large SCC, we conducted the aforementioned 3 experiments:

1. How many degrees of separation are there between two randomly chosen people? **4.788 (an average between 40,000 randomly chosen pairs) with a standard deviation of 1.103** See Figure 3 for distribution data.

2. What does the distribution of friendship look like? See Figure 2. Although this goes slightly against our intuition that many people would have many friends, it agrees with others' research findings [19][12] by having a power-law type distribution. Our past justification for the overwhelming # of few friendships suffices. Note that the top 10% of most popular users have at least 125 Los Angeles friends. Conversely, the least popular 10% of users have an average of 1.32 friends.

3. General statistics concerning structure: **average # of friends is 28.85.** When looking at all Los Angeles users' listing of friends, **only 21.45% of these friends belong to the Los Angeles network!** So, it appears that **Los Angeles is a very well-traveled, well-networked group of people, for the majority of their friends do not belong to the Los Angeles network!** One possible explanation is that since Facebook is expanding westward, maybe this has influenced their demographics of Facebook friends–as they may typically be from the East. Given our finding that Los Angeles is in fact largely consumed by a SCC, it should be known that the majority of users can reach anyone else within the network. At the end of this, we have those who have few friends and/or those who hide their listings of friends. For a visual of 100 typical users, please see Figure 4.

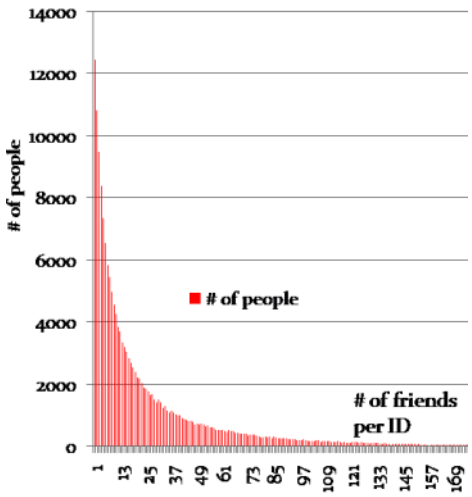

Figure 2: Distribution of users who have N friendships

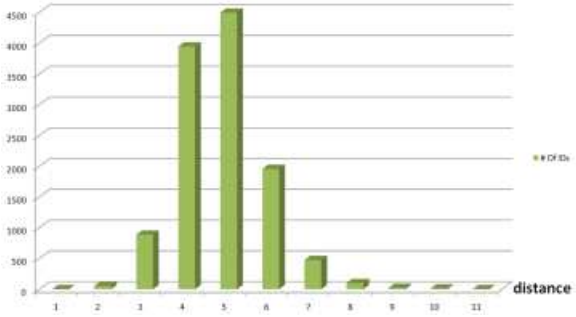

Figure 3: Distribution of degrees of separation between randomly chosen users

7. RELATED AND FUTURE WORK

We briefly touched on some various ways that others are making use of social networking data, from filtering e-mail spam to personalizing search results [10][3][15]. In addition, others have studied and attempted to model the evolution and dynamic growth of social networks, which we find particularly interesting [20][14]. If one looks at relationships on a small, individual scale, it can be rather trivial. How-

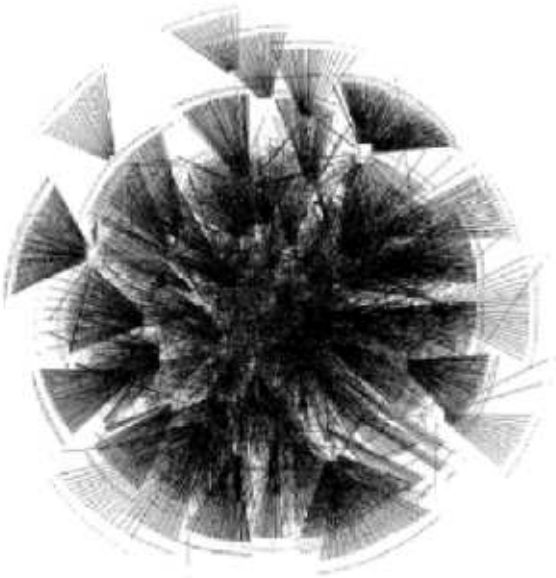

Figure 4: Distribution of degrees of separation between randomly chosen users

ever, on a large-scale, identifying large shifts in relationships could help in characterizing history.

As for future ideas, we would like to further explore the notion of influence. As we mentioned, a naive approach would be to implement a PageRank-like system. However, we feel it would be worthwhile to consider content and # of shared interest as a way of measuring how influential one is. Also, as for our structure, we make no distinction between our edges that connect friends. They are unweighted. However, if we weight the edges with values that correspond to their influence, or to some other metric about which we are concerned, then it would allow for many more opportunities of analyzing possible clusters of users.

With regard to users' interests, it is tempting to explore various NLP projects [9]. Also, it would be rather simple to extend our work to allow for users to find the corresponding users who most match them. This is less-research based, but would serve a somewhat useful, novelty tool. Currently, tools like this do not exist as social applications because it would obviously require some system to know the interests of all other users. Additionally, we would like to explore extraction of photo and video tags, along with other items that require deliberate efforts from the user to post. We feel that this strongly represents elements that users are interested in, possibly moreso than the often generic, ill-formed *interests* field.

Obviously, we would like to crawl more regional networks. Our only limitation is the immense time required to setup numerous Facebook accounts, along with required effort to babysit each crawling computer to ensure everything goes smoothly.

8. CONCLUSIONS

We attempted to better understand social networks, as they are becoming overwhelmingly popular. To gain an understanding, we first had to develop a robust, highly customizable crawler. With this, we were able to crawl a large percentage of the useful connections within the Los Angeles network. We were able to confirm that, in fact, Los Angeles, is very well connected: despite the fact that only 21% of Los Angeles users' friends belong to the network, there still exists a large SCC (176,000 users) whereby any user within it can reach any other user in roughly 4.8 friendship hops! We confirmed that the distribution of friendships agreed with others' research findings, as it can be characterized by a power-law distribution. In addition, we concluded that friends typically have almost twice as much in common than do randomly chosen strangers. These findings are worthwhile to note, but we have only begun to investigate the myriad opportunities that lie within the highly rich-in-content haven of social networks.

9. ACKNOWLEDGMENTS

This project would not have been possible without the help of others. Specifically, we thank our Professor Junghoo Cho, for he taught and exposed us to much regarding Web Information Management. Additionally, he provided web storage for us to archive all of our crawled data. Chris' Florida Tech colleague Tim Coulter was a tremendous help by volunteering to run our crawler non-stopped on his two personal computers. Likewise, UCLA colleague Mike Wilson was gracious enough to donate his machine to our crawling efforts. Last, we would like to thank the UCLA graduate students who spent their Thanksgiving vacation with their families, for it allowed us to fully utilize numerous computers in the abandoned graduate student lab.

10. REFERENCES

[1] Alexa - gloabl top 500 sites. October 30, 2007 http://www.alexa.com/site/ds/top500.php.
[2] Facebook user prevalence by geographic region. http://www.geocommons.com.
[3] Google co-op. http://www.google.com/coop/.
[4] Myspace hits 100 million accounts. August 9, 2006 http://mashable.com/2006/08/09/myspace-hits-100-million-accounts/.
[5] Myspace is the number one website in the united states according to hitwise. HitWise Press Release, July, 11, 2006 http://www.hitwise.com/press-center/hitwiseHS2004/social-networking-june-2006.php.
[6] Myspace reaches 190 million users. July 2007 http://www.myspace.com/.
[7] New face-book ads use your actions to sell ads. November 6, 2007 http://www.pcmag.com/article2/0,1759,2213009,00.asp.
[8] U.s. census bureau - los angeles (city) quick facts. http://quickfacts.census.gov/qfd/states/06/0644000.html.
[9] Wordnet. http://wordnet.princeton.edu/.
[10] Yahoo! my web. http://myweb2.search.yahoo.com.
[11] S. Brin and L. Page. Anatomy of a large-scale hypertextual web search engine. Proc. 7th International World Wide Web Conference, 1998.
[12] A. Broder, R. Kumar, F. Maghoul, and et al. Graph structure in the web. *Proceedings of the 9th*

international World Wide Web conference on Computer networks, 2000.

[13] P. S. Dodds. An experimental study of search in global social networks. *Science*, 2003.

[14] T. Fenner, M. Levene, G. Loizou, and et al. A stochastic evolutionary growth model for social networks. *Computer Networks*, 2007.

[15] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazi'res, and H. Yu. Re: Reliable email. In *In Proceedings of the 3rd Symposium on Networked Systems Design and Implementation*. NSDI, May 2006.

[16] S. H. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review*, 2006.

[17] S. Milgram. The small world problem. *Psychology Today*, 1967.

[18] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting social networks for internet search. In *In Proceedings of the 5th Workshop on Hot Topics in Networks*. HotNets, Nov 2006.

[19] A. Mislove, M. Marcon, K. Gummadi, P. Drushcel, and et al. Measurement and analysis of online social networks. *IMC Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[20] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *ACM*, 2006.

[21] D. J. Watts. Six degrees: The science of a connected age. *W. W. Norton*, 2003.