The Value of Being Linked In

Yaakov (J) Stein

April 2009

Abstract

I semi-empirically study the social networking sites such as *LinkedIn*. Such sites enable users to maintain contact information of people they know and trust (their *first degree connections* or *friends*), and to discover the friends of their friends (their *second degree connections*), and to access the friends of the friends of their friends (their *third degree connections*). Connections up to some degree (e.g., third) make up a *network*. I find the size of such a tree network grows sublinearly with time, even when its owner actively seeks out new friends. Under simplistic assumptions I find that the *value* of such a network to its owner is three times that of a standard contact list (containing only first degree connection). The total value of a network of N connections up to d degrees of separation to all its members scales as $N^{1+\frac{1}{d}}$. This is less than Metcalfe's law that states that the value of a fully connected network scales as N^2, but more than Odlyzko's law where the scaling is only $N \log N$. On the other hand, the cost of maintaining a large network increases faster than the value, and thus there is an optimal network size from the value point of view. Using models I am able to estimate the average degrees of separation between members of my network, and the size of the strongly connected cluster to which I belong.

1 Introduction

In early 2008 I started receiving from acquaintances inviting me to join their *networks* on various professional social networking sites, the most popular being LinkedIn [9], which at the time advertised over 25 million users (and is now quoting over 35 million). At first I ignored all such invitations, assuming that most have come either from people seeking jobs or from marketing types using these sites instead of business cards. However, I was subsequently led to reconsider my behavior. While tracking down prior art for a patent application I came across a reference to someone whom I had never met, and whom I could not locate via a general Internet search. A colleague performed a search on LinkedIn and promptly produced the desired contact information.

Acknowledgements : Thanks to all my connections who provided information and suffered my endless requests for assistance. Special thanks to Jonathan Levy for his help in programming the Web-scraping tool.

This anecdotal evidence lead me to ask whether it was possible to quantify the value of such large social networks. Does their value result from the fact that the tool accesses a large network or purely from the specific capabilities of the software platform? How much more valuable is a network of contacts than a conventional contact list, such as provided by conventional email clients?

It is evident that the social networking platform itself has some advantages, such as the fact that people update their own contact information. Thus, when your contact changes email address or telephone number, you need not manually update this information, or even know of the change until you need to contact him or her. However, securely updating contact information is surely a solvable problem, and one that probably should not require large interconnected networks.

So I was left with the question as to whether true value is derived from the fact that in addition to enabling access to your friends (called *contacts* in most conventional tools but *connections* in LinkedIn terminology) the user gains visibility to the friend's friends (second degree connections), and somewhat more limited knowledge of the friends's friends's friends (third degree connections). This entire network of first, second, and third degree connections is *much* larger than the set of friends (usually by a factor of between 1000 to 10,000), and has recently been the subject of interest. In order to process information in such networks, various machine readable formats for describing a FOAF (Friend Of A Friend) have been developed. Collections of FOAF entries have been called *social graphs* [7, 6], and Tim Berners Lee has called the set of all such data the *giant global graph* (GGG) [3]. Lee contends that the *value* of the GGG exceeds that of the WWW.

An extremely large social network of a different kind was recently studied by Leskovec and Horvitz [8]. They analyzed a month of communications activities of 240 million users of the Microsoft Messenger instant messaging system. They were able to construct the graph describing users who communicated, and found it to be richly connected, with an average of less than seven degrees of separation between randomly chosen users.

In order to better understand the advantages of a social network, I decided to join LinkedIn. I chose to join this particular network for several reasons.

- it is widely used by professionals in my areas of interest,
- unlike Facebook, its use is not blocked by my company's firewall,
- there are no undesirable instant messaging or similar *features.*

People who heard of my decision, gave me the benefit of their (anecdotal) experience. In particular I was told

- the size of the network increases "exponentially" (over time?)
- I would be able to build a network with a million connections with two weeks,
- I would find this network an extremely valuable tool,
- acquiring and maintaining such a network entails no cost,
- I would discover that any two people in a given field of interest are separated by no more than two or three degrees of separation,
- all 25 million LinkedIn users are somehow connected.

I decided to investigate these ideas by carefully keeping track of my network, as will be explained in section 2. The growth of the network, both analytically and empirically, is found to be governed by a power law, and not exponential. It took me over six weeks to reach a network size of one million.

In addition to providing a compendium of information on your friends that you can access from anywhere, the LinkedIn platform provides various tools to make contact with people in your network, such as introduction by a friend in common. Such access can be of value when trying to find potential collaborators or customers, when looking for a new job or looking for a candidate for a job you need to fill, etc. However, it is clear that not all network members are equally valuable, and thus the value of such a network only scales weakly with the total network size. In fact, the value of the network to me turns out to scale linearly in the number of friends (first degree connections), and not in the network size. I discuss *value* in section 3.

On the other hand, I questioned the idea that growing and maintaining a valuable network are completely free. Even if LinkedIn does not directly charge for maintaining a network, one certainly has to devote some time and effort to its upkeep. I shall show that the cost of maintaining a network scales quadratically in the number of friends, and thus must eventually exceed the linearly scaling value. In section 4 I empirically conclude that the trade-off is optimal at about $F = 500$ friends.

A key feature of social networks is that we can define the *degrees of separation* between any two network members. This represents the number of hops that need to be traversed between the two. Social network sites usually define the network of a particular user as the set of all users separated no more than some degree (e.g., 3). Thus although network membership is symmetric (if you are in my network then I must be in yours) it is not transitive (the friend of a third degree connection of mine is not necessarily in my network). I often fall into a self-centered way of thinking wherein my network is a tree emanating from me and extended to my third degree connections. In fact there are actually many connections between members of my network that have nothing to do with me (for example, my friends often directly know each other as well).

When finding a nonfriend using LinkedIn's search tools, the degrees of separation as well as partial information as to the connections along the path of separation are displayed. Likewise, when I view a friend's LinkedIn connections, friends in common are displayed first. However, LinkedIn does not make available the information necessary to directly deduce the degrees of separation between any two of my connections. In section 5 I present indirect inferences regarding connectivity and degrees of separation based on a pure tree model. I extend these results to somewhat more general models in section 6.

Finally, I build a detailed model of how a networks grows in section 7 and compare the analytic results to simulation and to the dynamics of my own network. From this model I was able to estimate the size of the connected cluster in which I reside. While LinkedIn claimed over 25 million users at the time, I seem to belong to a strongly connected cluster of only a few million people.

2 Linking In

In order to study social networks such as LinkedIn, I conducted an experiment over a period of several months, from the end of August 2008. I was interested in understanding the growth dynamics of such networks, the connection between the number of friends and the total network size, and in determining the value of maintaining such a network.

It is important to note that I studied *my* LinkedIn network, not *the* LinkedIn network. The latter is a complex graph with LinkedIn users as nodes and bidirectional edges between users who are friends. The only thing I know about it is the number of users (rounded to the closest five million). It would be interesting to know the size of the largest connected cluster, the distribution of the number of friends per user, the degrees of separation between users, etc. However, gathering such information would require access to LinkedIn's database. My information was limited to the information that LinkedIn provides to its users.

On the other hand *my* LinkedIn network is not simply a subgraph of this graph. In my network there is an order relation, assigning to every node its degree (between zero and three). Furthermore, there is an underlying tree structure, with each node being connected to me along a particular path or paths (see Figure 5).

A colleague who was aware of the experiment predicted that I would reach a network size of over one million within 2 weeks. It actually took much longer. After one week I had over 100 friends, and a network of close to 800,000 connections. But from then on the growth slowed considerably, and the million mark was passed only after over six weeks.

I started my network by accepting several outstanding invitations to join LinkedIn networks. Once accepted into someone's network I gained access to their contact list (although you can opt to keep your contact list confidential, in practice I found that very few people block access). I then scanned each contact list looking for people I wished to add to my network, and sent invitations accordingly. I never returned to rescan a friend's contact list afterwards. In a short time I started receiving, and accepting, numerous unsolicited invitations as well. If the need arose, I also used the search tools provided by LinkedIn to track down particular people I needed to contact.

In order to build a *valuable* network, I only invited, or accepted invitations from, people who met two criteria. The criteria were essentially the same ones that I normally consider before opening a contact record in my email client.

First, I had to actually *know* the person in question. The standard I used here is stronger than simply recognizing someone's name. I require that either I have met the potential friend on multiple occasions, or if on only on a single occasion, that we had conversed for at least one hour.

Second, there had to be some finite probability that I would need to contact the potential contact in the future. Such decisions were typically clear cut, but when in doubt I preferred to err on the side of letting in a possibly low-value contact rather than blocking someone whose contact information would be difficult to acquire later if I did end up needing it.

Note that I emphatically did *not* use profession as a criterion. Al-

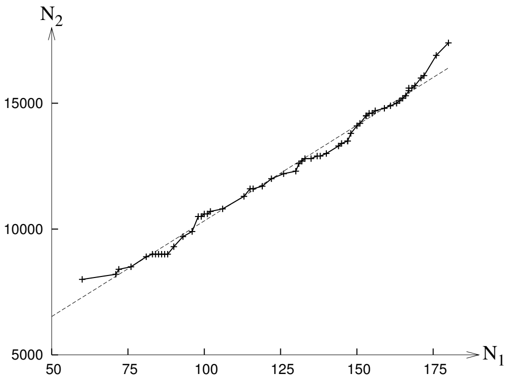

Figure 1: The number of second degree connections N_2 as a function of the number of friends (first order connections) N_1, for N_1 between 60 and 180 friends. For reference we superpose a line with slope 76 second order connections for each first order one.

though over 95% of my network naturally turned out to belong to the same high-tech and scientific communities as I do, I did not want to purposely block the outliers that may provide interesting connectivity between diverse people.

After a period of erratic initial network growth, I kept careful track of of the number of connections of degrees one, two, and three, and the size of the total network. In this fashion I could observe how adding a new friend affected the size of the network. Unfortunately, not all network growth is from my adding new friends, some is organic growth resulting from my friends adding new friends. So after amassing 60 contacts, I attempted for a while to grow my network as quickly as possible. In this way I could assume that the major contribution to network growth was from my activity, rather than from organic growth. I stopped this phase somewhat after passing the goal of a network of one million.

The next phase was an interval of six weeks in which I did not actively solicit addition of new friends, in order to gauge the rate of organic growth. In practice, a few new friends appeared as a result of invitations I had previously sent that were belated accepted.

In Figure 1 we see the number of second degree connections as a function of the number of first degree connections (friends). The points fall nicely on a line of slope 76, meaning that most of my friends had 76 friends of their own who were not on my list. The linear regression line does not intersect the origin, but rather crosses the y axis at the point where zero friends corresponds to 2724 second degree connections.

$$N_2 = 76N_1 + 2724$$

This is a clear indication that while we observe a straight line for this portion of the graph, the slope is decreasing with increasing number of friends. We will study the form of this graph in section 7.

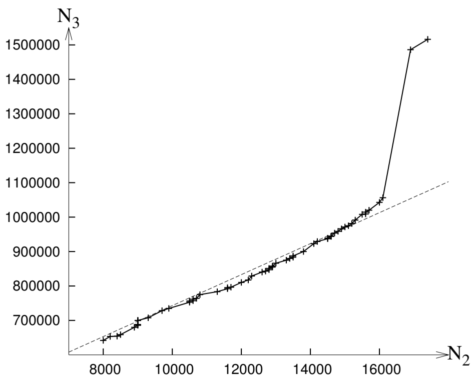

Figure 2: The number of third order connections N_3 as a function of the second order ones N_2, for N_1 between 60 and 180 friends. For reference we superpose a line with slope 46 third order connections for each second order one. Note the jump corresponding to the second phase of the experiment between 172 and 176 friends.

In order to ascertain how many new connections each second order connection adds to the network, I plot in Figure 2 the number of third degree connections as a function of the number of second order connections. For the first phase of the experiment the points fall discernibly, but somewhat noisily, on a line of slope 46,

$$N_3 = 46N_2 + 268729$$

meaning that most of my second order connections had 46 friends that were neither my friends nor previous second or third order connections of mine. The jump discontinuity at $N_1 = 172$ corresponds to the second phase of the experiment, where few new friends were added. This jump was noticeable but not very large in the previous graph, meaning that the organic network growth was mainly from second degree (and third degree) contributions.

To grasp the *big* picture, I plot the total network size N as a function of the number of friends N_1 in Figure 3. The network size is approximately linear in the number of friends, with about 3574 connections per friend.

$$N = 3574N_1 + 397159$$

Although this is a rather large slope, if I collect friends at a constant rate of n friends per unit time (so that $N_1(t) = nt + n_0$) this will result in the network size increasing only linearly over time.

$$\frac{\Delta N(t)}{\Delta t} = \alpha N_1(t) + A = \alpha\, n\, t + A'$$

Once I exhaust the obvious potential friends and my adding of new friends slows, (i.e., n decreases with time) the network increase will become sub-

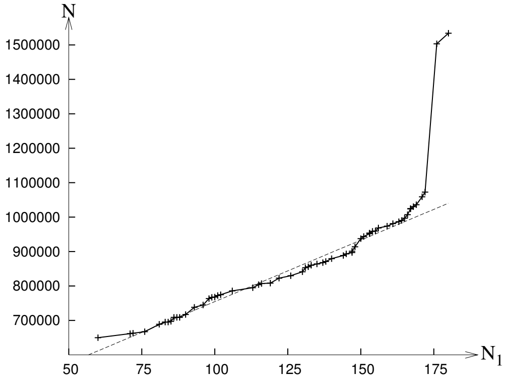

Figure 3: The size of the network N as a function of the number of first order connections N_1, for N_1 between 60 and 180 friends. For reference we superpose a line with slope 3574 network connections for each friend. Note the jump corresponding to the second phase of the experiment.

linear. This sublinear behavior is a far cry from the predictions of *exponential* growth.

It may be objected that this (sub)linear tendency is only characteristic of the first phase of the experiment, when organic growth can be neglected. The dramatic jump corresponding to the second phase demonstrates that organic growth can be significant, and may possibly contribute to superlinear growth rates. However, the dramatic jump in this depiction is misleading. The first and second phases were chosen to be approximately equal in time duration. During the first phase the number of friends increased from 0 to 170, and the network size from 0 to a million. In the second phase the number of friends barely increased, but the network increased to about a million and a half. So, in the same time duration, the organic growth of an already large network contributes only about half the growth rate of active expansion up to that point.

The organic growth rate is driven by first and second degree connections adding connections of their own. It thus is expected to have contributions proportional to N_1 and to N_2

$$\frac{\Delta N(t)}{\Delta t} = \beta N_1(t) + \gamma N_2(t) + B = \beta' N_1(t) + B'$$

where we have taken the linearity of N_2 in N_1 into account. Although in our second phase we did not actively increase N_1 in order to isolate the contributions, were we to continue increasing N_1 by n friends per unit time, we could substitute $N_1(t) = nt + n_0$ and and would see linearly increasing organic growth.

$$\frac{\Delta N(t)}{\Delta t} = \beta' nt + B''$$

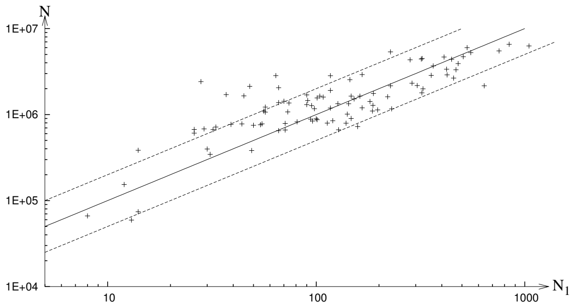

Figure 4: The size of the network N as a function of the number of first order connections N_1, for other people. Note that the data lie along a line with slope about 10,000 with some notable outliers.

Adding the active and organic contributions gives

$$\frac{\Delta N(t)}{\Delta t} = \alpha nt + A' + \beta' nt + B'' = \alpha' nt + A''$$

which is still only linear in time. Once n starts to decrease the total growth becomes sublinear.

Of course the precise growth dynamics will vary from network to network. While discussing with some of those on my contacts list, I discovered that some had much larger network sizes N for similar contact list sizes $F = N_1$. As one friend of mine put it, 'you are more connected than I am, but my connections are more connected than yours'.

To investigate how different the relationship could really be, I asked a number of friends to send me their 4-tuple (N_1, N_2, N_3, N) as presented on LinkedIn's *network statistics* page. Although $N = N_1 + N_2 + N_3$ and thus there are only three independent pieces of information here, it was worthwhile asking for all four numbers for error correction purposes, as friends would often copy the data incorrectly.

In Figure 4 I present N as a function of N_1 for other people's networks. Indeed variability is seen, but mostly for small N_1. For larger N_1 the data seems to fall along a line with slope of about 10,000 (the dashed lines represent slopes of 5000 and 20,000). The most significant outliers belongs to people unusual in my context, namely sales and marketing professionals. These people professionally meet a wide variety of people and tend to send out LinkedIn invitations to everyone they meet. Their contact list are rich in people with huge contact lists of their own (a fact substantiated by large N_2/N_1 ratios).

The arithmetic average of N_1 was 528, and of N_2 close to 67,000. However, these averages were severely distorted by the long tails of these distributions. A more representative statistic is the typical value (geometric mean), which for N_1 was about 90 and for N_2 about 8900. The first value is higher than the typical value in my own network of N_2/N_1, namely 83; but the second is lower than the typical value of $N3/N1$, which is close to 10,000.

3 Quantifying the Value of Contact Lists

Anyone who has ever had his list of emails or telephone numbers accidentally erased will attest to the value of such lists. Calling the number of entries in the contact list F, the quantitative value V is proportional to F. This trivially derives from the fact that if the contact information of any specific contact is lost or becomes corrupted, the list owner will be forced to spend time and effort, or alternatively would be willing to pay money, to restore this information. The cost that the list owner would be willing to incur to restore a contact's information is equivalent to the value assessed by the list owner to the possession of that contact information. If each of the F entries has value v then the total cost is $V = Fv$, which scales linearly with F.

This argument can be readily extended to the case where all contacts are valuable to some degree (else, the contact should be removed from the list) but not necessarily equally valuable. For such as case each contact has value v_i, the total value of the list is $V = \sum v_i = F < v >$ where $< v >$ is the average value. Thus, once again, the value of the list scales linearly with F.

In the LinkedIn social network, in addition to full access to $N_1 = F$ first degree connections, one can retrieve more limited information regarding N_2 second degree connections, and have some access to N_3 third degree connections.

Metcalfe's law [10] states that the value of a telecommunications network with N participants is proportional to N^2, i.e., $V \sim N^2$. (I use $V \sim N^2$ to indicate scaling, i.e., $V = O(N^2)$.) The reasoning behind this rule is simple to comprehend. Metcalfe is considering the total value of the network, i.e., the sum of the value to all participants. Since each participant can converse with N others, the value of the network to each scales as N, and the total value for N participants scales as N^2.

Of course Metcalfe's law makes the implicit assumption that as N grows all the new possible pairings remain valuable. In real life, as a network grows, its geographic extent and variability increases, reducing the value to an individual of each new participant in a manner that decreases with N.

Metcalfe's rule is not the most optimistic valuation for a network of N participants. Reed's law [13] goes even further based on the premise that in a network of N participants not only are N^2 conversations possible, but 2^N possible *conference calls* or *email lists*. Thus, according to Reed's law, $V \sim 2^N$.

Andy Odlyzko and Ben Tilly [12] dispute these strongly increasing valuations and propose a much weaker $N \log N$ behavior. They reason that the value's dependence on N must be superlinear, but not too strongly so. For any superlinear dependence, the value of two separate networks is less than a single unified one, causing strong pressure for disparate networks to be joined. For example, two networks of size N are separately worth $2N^2$, but together $(2N)^2$ – twice as much. This joining is indeed seen in the Internet, but many networks take too long time to merge for the N^2 law to be reasonable.

The simplest argument for $N \log N$ scaling is to assume that randomly

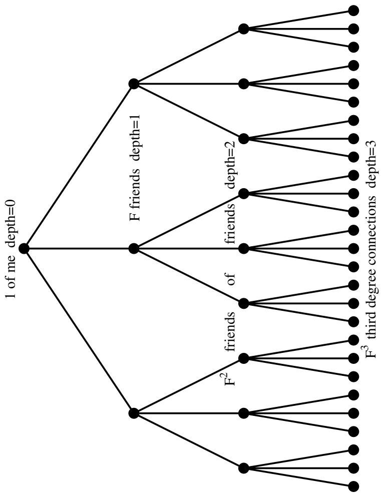
1 of me depth=0
F friends depth=1
F^2 friends of friends depth=2
F^3 third degree connections depth=3

Figure 5: Simple tree model.

chosen connections can be described by Zipf's law. One way of explaining Zipf's law is that for many naturally occurring sets of elements, if we sort the set by decreasing value, the 2^{nd} element will be approximately half as valuable as the first, the 3^{rd} element approximately one third as valuable, and the value of the k^{th} element will only be about $\frac{1}{k}$ of the first. Since the harmonic sum diverges logarithmically, the value of the network to each participant scales as $\log N$, and the total value as $V \sim N \log N$.

For a LinkedIn network, the value contributed by friends, i.e. first degree connections, must scale linearly in N_1, since these contacts are chosen by the user, and are thus valuable to at least some degree. However, the value of second degree connections is already much less, as your friend's brother, the hair-stylist, who lives half-way around the world, is of little interest to you.

In section 7 I will discuss how the size of the network N depends on the number of friends. For now, let us assume a simple *tree model* with *fan-out* F; wherein I have $N_1 = F$ contacts, each of whom contributes F distinct members to my network so that $N_2 = F^2$, and finally each of them contributes a further F new members, so that $N_3 = F^3$ (as in Figure 5). For large F we can assume that $N = 1 + N_1 + N_2 + N_3 \approx N_3 = F^3$

However, as mentioned above, not all members of my network contribute value, and thus we expect that the value of the network to me will scale more weakly than $V \sim N \sim F^3$. The question is how many second and third degree connections are valuable additions to the network. This question must be answered empirically.

When my contact list reached 100 members I performed an exhaustive study of my second degree connections. I went through the lists of all the non-shared connections of my connections (with the exception of three contacts who blocked this information). While the decision as to the value of a second degree connection is necessarily subjective, in practice the decisions were typically easily made. I rated as valuable a connection that fulfilled at least one of the following criteria:

- I immediately recognized the connection's name
- the connection and I have previously communicated more than once
- the connection and I share at least two areas of interest.

Note that these criteria are more lenient than those I use to accept someone into my network as a first degree connection.

In this search I found about 120 valuable second degree connections. More interestingly, I almost always found one or two valuable second degree connections in the list of each friend. Perhaps surprisingly, the number of valuable connections was not greater for first degree connections with more connections than for those with fewer (i.e., larger lists tended to have more chaff). Thus the number of *valuable* second degree connections M_2 was only slightly larger than the number of first degree connections $M_1 = N_1 = F$.

Let's assume that each first-degree connection contributes exactly m valuable second-degree connections. Then the number of second-degree connections is $M_2 = mF$. Although not directly verifiable, it would seem to be highly improbable for a non-valuable second-degree connection to contribute a valuable third degree connection. On the other hand, valuable second degree connections, although held to a looser standard than friends, are not really that different from my first-degree connections. It would thus seem probable for each of them to contribute about m valuable connections to the network.

Taking these two assumptions to be true, we deduce that $M_3 = m(mF) = m^2F$, and

$$M = M_1 + M_2 + M_3 = F + mF + m^2F = (1 + m + m^2)F = f_3(m)F$$

where,

$$f_d(m) = \sum_{p=0}^{d-1} m^p$$

for example,

$$f_3(m) = \sum_{p=0}^{2} m^p = \begin{cases} 1 & m = 0 \\ 3 & m = 1 \\ 7 & m = 2 \\ 13 & m = 3 \end{cases}$$

and the value of the network to me scales like $f(m)F = f(m)N_1$. In our simple tree model, $N \sim N_1^3$ so that my valuation of my network is $V \sim N^{\frac{1}{3}}$. This result can be readily extended to a network with $d > 3$ degrees of freedom, where my valuation of my network would scale as $V \sim f_d(m)F \sim N^{\frac{1}{d}}$.

Of course, this is my valuation of the network, that is, the value of the network to *me*. On the other hand, members of my network find value in their friends, and some value to all their connections including myself. Assuming that their situation is not too different from mine, each of the N members of my network also finds value $V \sim N^{\frac{1}{d}}$ in their respective networks. hence, the total value of the network for all N members is N times the value for each, which thus scales as $N^{1+\frac{1}{d}}$.

For simple contact lists $d = 1$ and thus the total network value $V \sim N^2$ as required by Metcalfe's law. For networks such as LinkedIn $d = 3$, so that $V \sim N^{\frac{4}{3}}$. This is weaker than Metcalfe's law $V \sim N^2$, but stronger than Odlyzko's law $V \sim N \log N$.

4 How Linked In Should You Be?

In the previous section we saw that the value to me of my network increases linearly in the number of friends, $V = \alpha F$. If the value of my network increases with its size, then why am I not motivated to grow my network without limit?

Although LinkedIn does not charge for using the basic service, there are nonetheless indirect costs to maintaining a large network. For example,

- any of my friends may ask me for an introduction to another of my friends; since the number of pairs of friends is (in the tree model with fan-out F) approximately F^2, the probability of this occurring is βF^2 for some small β,
- second degree connections who discover me on the connection list of a friend in common may ask me to join their network. This frequently happens even if this second degree connection does not meet the criteria I use for adding new friends them to my network. Since the number of second degree connections is F^2 (and while I only consider mF of them valuable additions to *my* network, any of the F^2 *may* consider me valuable additions to *their* network), the probability of this happening is γF^2 for some small γ.

Although the probabilities of each of these events is small, it takes me some time and effort to respond to them, and their cost to me is proportional to their combined probability $(\beta + \gamma)F^2$, which, being quadratic in F, must at some point pass the linearly increasing value $V = \alpha F$.

Unfortunately, there does not seem to be any direct way of directly estimating the constants α, β and γ. In order to better understand the tradeoffs involved I approached my friends and asked them how useful LinkedIn was to *them* as a tool, and how much effort they expend on its upkeep. I sorted the responses according to the size of the friends network.

I did not bother asking connections with fewer than 50 connections, since these were deemed to be casual users who do not put significant effort into their network, nor would they be likely to see it as very valuable.

The group of friends with 50-200 friends of their own almost unanimously reported that the value of their networks far exceeds its any maintenance costs. In fact, almost all were quite surprised to hear me mention any costs at all.

On the other hand, people with huge networks of over 700 friends almost all reported that the upkeep costs were excessive. Several remarked that they were no longer actively updating their account information.

Some typical responses were :

- people are queued up waiting for me to accept their invitations,
- I haven't been taking care of my LinkedIn network, I am too busy with Facebook and MySpace,
- sorry for taking so long to accept your invitation, I get so many that I only process them once a month,
- too many people are asking me to join to get access to my contacts,
- I have to do something for someone twice a week, but only use it myself every few months,
- I have stopped accepting invitations - there are too many job seekers out there.

People with between 200 and 700 friends seem to divided about half and half. Some put a lot of time into keeping building their profile and keeping it up to date, while others cut and paste something once and have since forgotten about it. When questioned specifically as to how often they retrieve useful information from the network versus how frequently they are asked to take some action for someone else, the numbers ranged from *twice a week* to *it's only happened a few times* for both questions. We can surmise that the optimal size is somewhere around $F = 500$. Under that, the upkeep costs are small and useful information can be retrieved relatively inexpensively. Over that, the maintenance costs become excessive, and people that well connected apparently have alternative ways of getting the needed information.

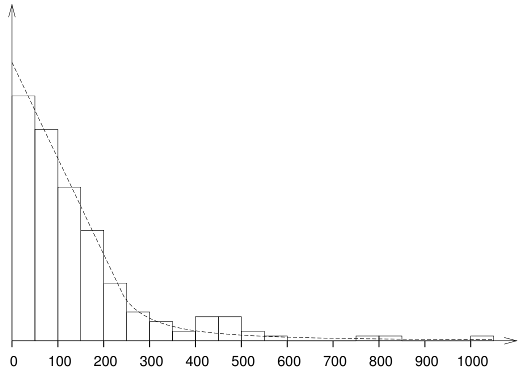

Figure 6: Histogram of the number of friends for each of my friends. For small numbers of friends the distribution decreases linearly in F, but for large numbers it follows a (long-tailed) power law.

Another way of testing our conclusion is to check the distribution of the number of friends for all my LinkedIn friends. This information is easily retrieved from my connections list for those friends with fewer than 500 contacts. For those with over 500 contacts LinkedIn states '500+' on that page, so I had to actually visit their contact pages. For those with block contact lists I needed to email them requesting their information.

Figure 6 depicts the histogram of number of friends for 190 of my friends. For friends with small to intermediate networks (less than 250 friends) the histogram is approximately as linearly decreasing. For larger networks the distribution is *long-tailed* (in the figure I somewhat arbitrarily use the F^{-2} power law). While there is insufficient data to be able to accurately determine the exponent of the power law, it is clear that the behavior is that of a *scale-free* network [2] rather than exponential decrease of a random network [4].

The remarkable feature here is the clear existence of delineated regimes. The linear decrease occurs in the realm where value exceeds cost, and thus people are motivated to continue adding friends to their networks. The power law decrease is in the regime where the cost exceeds value. In between these two is a noticeable (and statistically significant) bump in the histogram corresponding to more than expected friends of mine having between 400 and 500 friends of their own. This overabundance comes at the expense of fewer than expected friends having contact lists with 550 to 750 contacts. The existence of the regimes and the bump lend credence to our conclusion that there is a cross-over somewhere in the neighborhood of 500 contacts.

5 Degrees of Separation

It is of great interest to know the number of degrees of separation between two members of my network. Milgram's *six degrees of separation* rule (also known as the *small world* phenomenon) [11] states that anyone is at most six degrees of separation from anyone else on the planet. Milgram did not invent the number six. The first to state six degree of separation concept was Frigyes Karinthy (1887-1938), the Hungarian author, playwright, poet, and journalist. Karinthy believed that the modern world was shrinking in the sense that technology was making social distances much smaller than physical distances. In a short story entitled *Chain-Links* in his 1929 volume of short stories *Everything is Different* his characters conjecture that any two individuals could be connected through at most five acquaintances.

While six degrees of separation has become popularized for any two humans, people working in the same narrow field of international interest are probably much closer than that. One well-known case is that of *Erdös numbers* [5] that have become entrenched in the mathematician's folklore. Paul Erdös (1913-1996) wrote over 1,400 mathematical research papers with over 500 co-authors. Erdös himself is given Erdös number 0, and his co-authors (including Odlyzko whose law was mentioned in section 3) have Erdös number 1. Co-authors of people with Erdös number 1 (who are not Erdös nor have written a joint paper with Erdös) have Erdös number 2, and so on recursively. Anyone not on the *Erdös collaboration graph* has an infinite Erdös number. My own Erdös number is apparently 5 through three different paths. Most active mathematicians have relatively low Erdös numbers (i.e., few degrees of separation from Paul Erdös) and are even closer to randomly chosen active mathematicians [1].

By construction, I am never further than three degrees of separation from anyone in my LinkedIn network of first, second, and third degree connections. However, two people in my network may be up to six degrees of separation from each other, three from the first person to me, and three back to the second person.

What is the average distance between any two members of my network? In a pure tree model with large F, the great majority of network members are third degree connections, and a randomly picked pair of them will mostly likely be separated by the full 6 degrees. However, for finite F the average separation will be smaller.

Let's define the average separation $< d >$ as the expectation of the separation when we pick two network members i and j at random. This will be the same as finding the expected number of degrees of separation $< d_i >$ of a randomly chosen member j from a given i, averaged over all possible i.

$$\begin{aligned} < d > &= \frac{1}{N(N-1)/2} \sum_{i \neq j} d_{ij} \\ &= \frac{1}{N(N-1)} \sum_{i<j} d_{ij} \\ &= \frac{1}{N} \sum_i \frac{1}{N-1} \sum_j d_{ij} \\ &= \frac{1}{N} \sum_i < d_i > \end{aligned}$$

Let us assume that the network is a tree with fan-out F. From symmetry it is clear that all i of the same depth k on the tree will have the same $< d_i >$. Thus rather than averaging over all i to find $< d >$, we need only perform the weighted average over all depths.

$$< d >= \frac{1}{F^3 + F^2 + F + 1} \sum_k N_k < d_k >$$

This significantly reduces the analysis effort required to find the desired average.

The degrees of separation between two nodes on our network tree depend on the depths of the two nodes, and the depth of their (borrowing terminology from family trees) Most Recent Common Ancestor (MRCA). In Table 1 I give all the information required to calculate the average degrees of separation between two nodes on a network tree. The first column contains the depth k of the node for which we want to calculate $< d_k >$. The second column lists the depth of the node to which we want to find the degrees of separation. In general there will be several rows for each such depth, differentiated by the depth of the MRCA. The third column gives the number of such second nodes, and the final column gives the degrees of separation. Note that each pair of nodes appears twice in the table.

In order to calculate $< d_k >$ we need to average over all rows belonging to a given first column, of the fourth column weighted by the third column. Next, we need to divide by the total number of nodes on the tree excepting the present one.

$$d_0 = \frac{1}{F^3 + F^2 + F} \left(3F^3 + 2F^2 + F \right)$$

from	**to**	**MRCA**	**number**	**separation**
0	1	0	F	1
	2	0	F^2	2
	3	0	F^3	3
1	0	0	1	1
	1	0	$F-1$	2
	2	1	F	1
		0	F^2-F	3
	3	1	F^2	2
		0	F^3-F^2	4
2	0	0	1	2
	1	1	1	1
		0	$F-1$	3
	2	1	$F-1$	2
		0	F^2-F	4
	3	2	F	1
		1	F^2-F	3
		0	F^3-F^2	5
3	0	0	1	3
	1	1	1	2
		0	$F-1$	4
	2	2	1	1
		1	$F-1$	3
		0	F^2-F	5
	3	2	$F-1$	2
		1	F^2-F	4
		0	F^3-F^2	6

Table 1: The degrees of separation **from** a node of given depth on a tree **to** another. The MRCA is the Most Recent Common Ancestor, i.e., the deepest node through which the path of separation passes.

$$\begin{aligned}
d_1 &= \frac{1}{F^3+F^2+F}\left(4F^3+F^2-1\right)\\
d_2 &= \frac{1}{F^3+F^2+F}\left(5F^3+2F^2-F-2\right)\\
d_3 &= \frac{1}{F^3+F^2+F}\left(6F^3+3F^2-3\right)
\end{aligned}$$

Finally, the average degrees of separation is the weighted average over all depths k of $< d_k >$.

$$\begin{aligned}
< d > &= \frac{1}{(F^3+F^2+F+1)}\left(d_0+Fd_1+F^2d_2+F^3d_3\right) \qquad (1)\\
&= \frac{1}{(F^3+F^2+F+1)(F^3+F^2+F)}\left(6F^6+8F^5+6F^4\right)
\end{aligned}$$

We depict this as a function of fan-out F in Figure 7.

For $F=1$, we have the value $\frac{5}{3}$. This is because for this case the node with depth 0 and the single node with depth 3 both have one node with each of the degrees of separation 1, 2, and 3, for an average of 2. While the

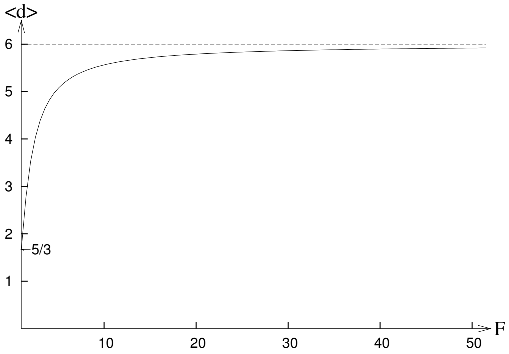

Figure 7: The average separation distance between two members of a tree network $< d >$ as a function of the fanout F.

2 nodes of depth 1 and 2 have two nodes with one degree of separation, and one with 2 degrees of separation, for an average of $\frac{4}{3}$. Thus the overall average is half of $2 + \frac{4}{3}$, i.e., $\frac{5}{3}$.

As expected, as $F \to \infty$ we see that the average degrees of separation approaches 6. This means that the pure tree model predicts that although my connections are all on LinkedIn, and although they are all no more than three degrees of separation from me, they are just about as far from each other as any two people on the planet.

6 Jumping Off the Tree

The above calculations assumed a pure tree network. This can not really be the case for my LinkedIn network. When I look at my contact's connections I almost always see first a list of shared connections, meaning that my connections are themselves connected. This means that there are *cross-links* that break the pure tree model, and I am led to ask whether I can predict the degrees of separation in a more general setting.

If the probability of two of my friends having a friend in common is not too large, then a perturbed tree model is reasonable. Such a perturbed structure is still based roughly on a tree with fanout F, but between nodes of depth 1 (my friends) the probability of a cross-link connection is p_1.

Before recalculating the average degrees of separation for such a model, I need to find p_1. I did this by *scraping* the LinkedIn web site, as described in Appendix A. I did this when my contact list consisted of 180 contacts, but I neglected those with blocked contact lists or with fewer than 30 contacts on their lists. This left me with 144 friends, of which I could check $144 * 143/2 = 10,2960$ pairs. 3,515 of these pairs of friends had friends common other than myself, giving very close to $p_1 = \frac{1}{3}$.

While putting in the effort of collecting this information, I decided to compute in addition the *overlap* between the contact lists of two friends. I define the overlap between two friends i and j, as the number of contacts

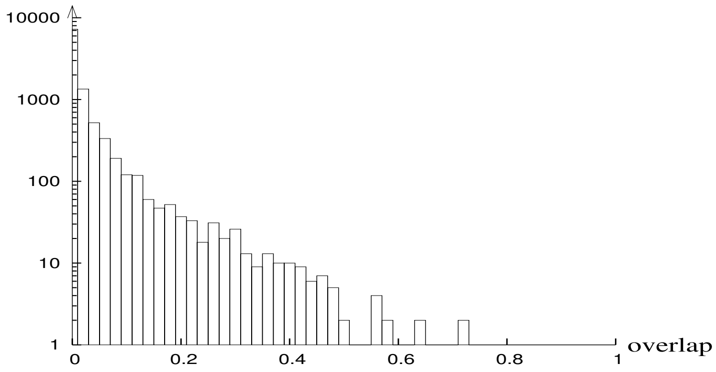

Figure 8: Logarithmic display of the histogram of overlap between members of my network.

they have in common (not including myself) F_{ij} divided by the smaller of the number of friends i has or j has (once again, not including myself).

$$O_{ij} = \frac{F_{ij}}{\min(F_i, F_j)}$$

A histogram of these overlaps is displayed on a logarithmic scale in Figure 8. The overlap decreases from two thirds of the pairs with zero overlap, to less than a fifth of a percent with overlap over $\frac{1}{2}$. There is a slight overtendency for overlaps to be around $\frac{1}{3}$.

The detailed information gathered by scraping my LinkedIn network can be used for further studies, such as blind determination of relationships between people. Anecdotally, I discovered that two friends of mine with a surprisingly high overlap had once worked together. It is simple to define a distance measure between two members of my network based on the overlap, and to perform clustering and cladistic analysis of this data. Such an analysis is in its early stages and will be described elsewhere.

Now we can reperform the calculation that lead to Equation 1 taking cross-link connections into account. The simplest topology that we can study is based on the pure tree model, but augmented with cross-links between friends with probability $p_1 = \frac{1}{3}$. This is the only model for which we have access to the probabilities, since one can not scrape from LinkedIn information regarding the probability of cross-links between higher degree connections. In addition, the probability of such higher degree cross-links will undoubtedly be much lower.

Repeating our computation is straightforward. The intermediate results are presented in Table 2 and the average degrees of separation from connections at the different depths are easy to find.

$$\begin{aligned} d_0 &= \frac{1}{F^3 + F^2 + F}\left(3F^3 + 2F^2 + F\right) \\ d_1 &= \frac{1}{F^3 + F^2 + F}\left((4 - p_1)F^3 + F^2 - (1 - p_1)\right) \end{aligned}$$

from	**to**	**MRCA**	**number**	**separation**
0	1	0	F	1
	2	0	F^2	2
	3	0	F^3	3
1	0	0	1	1
	1	x	$p_1(F-1)$	1
	1	0	$(1-p_1)(F-1)$	2
	2	1	F	1
		x	$p_1(F^2-F)$	2
		0	$(1-p_1)(F^2-F)$	3
	3	1	F^2	2
		x	$p_1(F^3-F^2)$	3
		0	$(1-p_1)(F^3-F^2)$	4
2	0	0	1	2
	1	1	1	1
		x	$p_1(F-1)$	2
		0	$(1-p_1)(F-1)$	3
	2	1	$F-1$	2
		x	$p_1(F^2-F)$	3
		0	$(1-p_1)(F^2-F)$	4
	3	2	F	1
		1	F^2-F	3
		x	$p_1(F^3-F^2)$	4
		0	$(1-p_1)(F^3-F^2)$	5
3	0	0	1	3
	1	1	1	2
		x	$p_1(F-1)$	3
		0	$(1-p_1)(F-1)$	4
	2	2	1	1
		1	$F-1$	3
		x	$p_1(F^2-F)$	4
		0	$(1-p_1)(F^2-F)$	5
	3	2	$F-1$	2
		1	F^2-F	4
		x	$p_1(F^3-F^2)$	5
		0	$(1-p_1)(F^3-F^2)$	6

Table 2: The degrees of separation **from** a node of given depth **to** another, when there is the probability of c_1 of *cross-links* between nodes of depth one. Paths passing through a cross-link are identified by an 'x' in the MRCA column.

$$
\begin{aligned}
d_2 &= \frac{1}{F^3+F^2+F}\big((5-p_1)F^3+2F^2-F-(2-p_1)\big) \\
d_3 &= \frac{1}{F^3+F^2+F}\big((6-p_1)F^3+3F^2-(3-p_1)\big)
\end{aligned}
$$

Finally, the average degrees of separation between any two connections is the weighted average over all depths k of $< d_k >$.

$$
< d >= \frac{\big((6-p_1)F^6+(8-p_1)F^5+(6-p_1)F^4+p_1F^2+p_1F\big)}{(F^3+F^2+F+1)(F^3+F^2+F)} \qquad (2)
$$

Note that for $p_1 = 0$ we return to equation 1 as required. We depict the

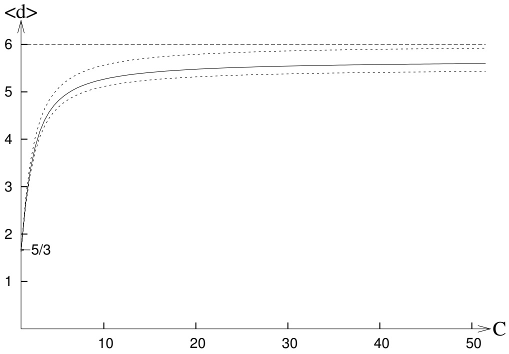

Figure 9: The average separation distance between two members of a perturbed tree network $< d >$ as a function of the fanout F for three probabilities of first degree cross-links, $p_1 = 0$ (unperturbed tree), $p_1 = \frac{1}{3}$ (the observed value in my network), and $p_1 = \frac{1}{2}$.

degrees of separation as a function of fan-out F in Figure 9 for $p_1 = 0$ (the pure tree model), $p_1 = \frac{1}{3}$ (the observed value for my network), and for $p_1 = \frac{1}{2}$ (the highest value for which it is still sensible to call the structure a perturbed tree).

As we can see, the degrees of separation indeed declined due to the cross-links, but not by very much. In fact, it is easy to deduce from equation 3 that as $F \to \infty$ the average degrees of separation approaches $6 - p_1$. This is not surprising as for large F, most of the connections in the network are at depth 3 and separated from each other by either 6 or 5 degrees of separation. The probability of the latter is p_1, so that $< d >= 6(1 - p_1) + 5p_1 = 6 - p_1$.

So the tree perturbed by cross-links between friends only has the potential of reducing the average degrees of separation of a large network from 6 to $5\frac{1}{2}$. The next step would be to allow cross-links between second degree connections, and between friends and second degree connections. While higher degree connections are less likely to know each other and thus have cross-links with lower probability, the cross-links that *do* exist have more potential of reducing $< d >$. In fact, if there are a sizeable number of cross-links between third degree connections, then $< d >$ will be close to 1.

We have no direct way of estimating the probability of there being cross-links of these types, but if p_1 is small enough so that our tree model is meaningful, arguments similar to those of Section 3 lead us to believe the probability of two second degree connections being friends is $p_2 = p_1^2$. The probability of cross-links spanning the first and second degrees would presumably be something in between.

Actually, we were somewhat hurried in concluding that the probability of a link between *any* two second degree connections would be p_1^2. Of the

$\frac{F^2(F^2-1)}{2}$ pairs of second degree connections, $F\frac{F(F-1)}{2}$ share MRCA of the first degree, and thus would be expected to have a cross-link with probability p_1! This leaves $\frac{F^3(F-1)}{2}$ pairs with probability p_1^2 so that the average probability is $p_2 = \frac{p_1+Fp_1^2}{F+1}$. So far large F it is indeed true that $p_2 \approx p_1^2$.

In similar fashion we expect for large F for the probability of cross-links between two third degree connections to be $p_3 = p_2^2 = p_1^4$. For large F we can neglect all other cross-links and all pairs of connections except those whom would have been at separation 6 in the pure tree model. These connections are directly connected, and thus at separation 1 with probability p_1^4, and otherwise at separation 6.

$$< d >= (1 - p_1^4) \cdot 6 + p_1^4 \cdot 1 = 6 - 5p_1^4 \tag{3}$$

For $p_1 = \frac{1}{3}$ this works out to be about 5.94, although it is lower for finite F.

So higher degree cross-links are even less effective at reducing $< d >$ than cross-links between friends. Even taking both effects together would not reduce the average value to less than 5. Perturbed tree models do not lead to low numbers of degrees of separation.

7 Growing the network

Leskovec and Horvitz [8], having access to the raw data of their network, were able to discover that 99.9% of the 180 million nodes belong to the largest connected component. In this section I'll attempt to indirectly deduce the size of the connected component in which my network resides, based on the very partial information retrievable from LinkedIn.

In the previous section we saw the consequences of deviating from a pure tree model. A more radical deviation from the tree model is a random IID bond model wherein we assume that there are N registered LinkedIn users, between every pair of which there is a link with some probability p. This means that on average each user has $p(N-1)$ friends, and that there are about $pN(N-1)/2$ friendship relationships in the network. A well-known theorem of Erdös and Rényi theorem [4] states that in large random graph of N nodes with probability p for edges between any two nodes, if the probability p exceeds $\frac{c}{N}$, then there is single connected component containing $O(N)$ of the nodes, and no other connected component has more than $O(\log N)$ connections. Translated into social network language this states that if the users choose friends at random, then if on the average each user has more than one friend, then there is a giant connected component with a finite percentage of all users, and all other connected components are local cliques uninterested in the rest of humanity. Furthermore, Erdös and Rényi proved that the typical degrees of separation between two users scales as the logarithm of N.

However, the assumption that any two LinkedIn users have the same probability of being friends is as unrealistic as the pure tree model. Barabasi's approach [2] is more appropriate as it describes dynamically growing networks, and preferential connectivity between nodes. It may be appropriate for description of the full LinkedIn network, but does not seem to properly capture the features of my *personal* network.

In this section I take a different approach. I incrementally grow my connection tree, assuming that there are a finite number of potential connections N from which to grow it. At the stage when I have amassed k friends I will call the total number of connections in my network $N(k)$, and the number of first, second, and third degree connections $N_1(k) = k$, $N_2(k)$, and $N_3(k)$ respectively. I call the number of second degree connections gained by adding the k^{th} friend $n_2(k)$, i.e., $N_2(k) = N_2(k-1) + n_2(k)$. Similarly, the finite difference of $N_3(k)$ is called $n_3(k)$, so that $N_3(k) = N_3(k-1) + n_3(k)$.

I assume that before I start growing my tree everyone else in the network has already chosen F friends, and call the probability of any any specific user having another user as a friend $p = \frac{F}{N}$.

I start growing my network by choosing a first connection. This gives me $N_1(1) = 1$ first degree connections, and $N_2(1) = n_2(1) = F$ second degree connections. When I choose my second connection (different from the first), I have $N_1(2) = 2$ first degree connections, but somewhat fewer than $2F$ second degree connections. In fact, the probability that any one of my second connection's connections was also chosen by my first connection is the ratio of his number of connections to the number of users in the network, i.e. $\frac{F}{N} = p$. Thus the probability that it was not one of his connections is $1-p$, and the expected number of *new* connections of the second degree is $n_2(2) = F(1-p)$. So the total number of connections of the second degree is $N_2(2) = N_2(1) + n_2(2) = F + F(1-p) = 2F - pF$, indeed slightly less than $2F$. I neglected the possibility that my first friend is also a friend of my second friend, but assuming $F >> 1$ this will not significantly change the results.

Now, adding a third (different) connection brings me to $n_1(3) = 3$, but the probability of overlap with either of the first two connections is $\frac{N_2(2)}{N}$, so that

$$\begin{aligned} n_2(3) &= F(1 - \frac{N_2(2)}{N}) \\ &= F\{1 - \frac{F}{N}(1-p)\} = F\{1 - p(1-p)\} \\ &= F(1 - 2p + p^2) \end{aligned}$$

and

$$\begin{aligned} N_2(3) &= N_2(2) + n_2(3) \\ &= F(2-p) + F(1 - 2p + p^2) \\ &= F(3 - 3p + p^2) \ . \end{aligned}$$

In like fashion

$$\begin{aligned} n_2(4) &= F(1 - 3p + 3p^2 - p^3) \qquad & N_2(4) &= F(4 - 6p + 4p^2 - p^3) \\ n_2(5) &= F(1 - 3p + 3p^2 - p^3) \qquad & N_2(4) &= F(4 - 6p + 4p^2 - p^3) \end{aligned}$$

and in general, the total number of second degree connections after adding

k connections is

$$\begin{aligned} N_2(k) &= F\left(k - \binom{k}{2}p + \binom{k}{3}p^2 - \ldots \pm p^{k-1}\right) \\ &= F\frac{1}{-p}\left(k(-p) + \binom{k}{2}(-p)^2 + \binom{k}{3}(-p)^3 + \ldots (-p)^k\right) \\ &= F\frac{1}{-p}\left(-1 + (1-p)^k\right) = F\frac{1-(1-p)^k}{p} \\ &= N\left(1-(1-p)^k\right) \end{aligned}$$

For small p and k this reduces to

$$N_2(k) = N\left(1-(1-kp)\right) = Nkp = Fk$$

which is linear in k with slope equal to F as expected. On the other hand, were k to grow to become very large, $N_2(k)$ would start leveling off, since the probability of second degree connections having friends in common becomes large.

We can apply this same technique to the transition from second degree connections to third degree ones. After I add my k^{th} friend, I have $N_2(k)$ second degree connections, each of whom picks F friends. However, because of the overlap, this results in fewer than $N_2(k)F$ new third degree connections. It is easy to see that

$$\begin{aligned} n_3(k) &= N_2(k)F\left(1 - \frac{N_1(k)+N_2(k)+N_3(k-1)}{N}\right) \\ &= N\left(1-(1-p)^k\right)F\left(1 - \frac{k + N\left(1-(1-p)^k\right) + N_3(k-1)}{N}\right) \end{aligned}$$

which taken along with $N_3(k) = N_3(k-1) + n_3(k)$ gives a rather complicated recursion for $N_3(k)$.

However, for large F we can neglect k and $N_2(k)$ as compared to $N_3(k)$,

$$n_3(k) = N_2(k)F\left(1 - \frac{N_3(k-1)}{N}\right)$$

so that $N_3(k)$ relates to $N_2(k)$ in the same way that $N_2(k)$ relates to $N_1(k) = k$.

In order to check the approximations, I simulated the behavior of a small network. My simulated environment consists of 65,536 users (each user is given a 16-bit identifier). I initialize an array of length 65,536 that represents the degree of each user in my tree by entering a suitably large number in each position. The fanout was chosen to be $F = 64$.

I choose friends one at a time by randomly selecting a 16-bit number, and entering the value 1 in the array of 65,536 users. Each of these friends randomly selects $F = 64$ friends (my second degree connections) by randomly choosing F 16-bit numbers, and entering the value 2 in the array, unless there is already a 1 there. Each of my second degree connections chooses $F = 64$ friends (my third degree connections) at

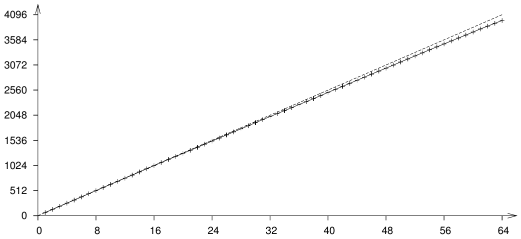

Figure 10: The results of network growth simulation. The graph depicts the number of second degree connections $N_2(k)$ as a function of the number of friends k, for k from 1 to 64. Note the slight deceleration of growth as compared to linear growth expected for the infinite network case.

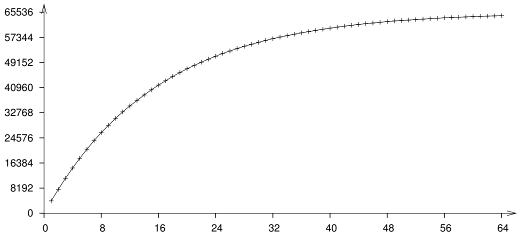

Figure 11: The results of simulating network growth. The graph depicts the total number of connections $N_1(k) + N_2(k) + N_3(k)$ as a function of the number of friends k, for k from 1 to 64. Note the strong leveling off of growth due to finite network size.

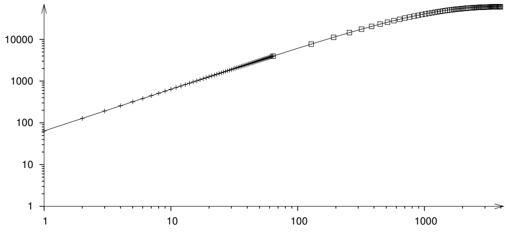

Figure 12: The results of simulating network growth. The graph depicts $N_2(k)$ vs. $N_1(k) = k$, and $N_3(k)$ vs. $N_2(k)$ on a logarithmic scale as compared to the expected curve.

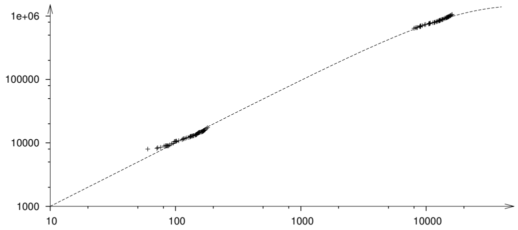

Figure 13: Finding N_{LI} by matching the predicted growth curve. The graph depicts $N_2(k)$ vs. $N_1(k) = k$ and $N_3(k)$ vs. $N_2(k)$ on a logarithmic scale as compared to the expected curve. In this graph $N_{LI} = 2,000,000$.

random and enters a 3 into the array, unless the value already in that position is less than 3. Finally, I scan the array counting how many 1s, 2s, and 3s are present. I repeat this procedure for between 1 and 64 friends.

Figure 10 depicts the number of second degree connections (i.e., number of entries in the array with value 2) as a function of k, for k between 1 and 64. Were there an infinite number of users we would expect this to be a linearly increasing function, with 64 new second degree connections for each new friend, and $64^2 = 4096$ second degree connections at the end. The actual behavior displays a slight deceleration of growth in comparison with the superposed straight line.

Figure 11 depicts the total network size (i.e., number of entries in the array with value less than or equal to 3) as a function of k, for k between 1 and 64. For an infinite number of users we would expect

$$N_3(k) = N_1(k) + N_2(k) + N_3(k) = k + kF + kF^2 = k(1 + F + F^2)$$

which for $k = 64$ would reach $N_3(64) = 64 + 64^2 + 64^3 = 266,304$. Obviously the effect of there only being 65,536 possible connections is very pronounced.

Figure 12 depicts both the number of second degree connections as a function of the number of first degree connections (from Figure 10) and the number of third degree connections as a function of the number of second degree ones. The similarity of behavior of $N_3(k)$ vs. $N_2(k)$ as compared to $N_2(k)$ vs. $N_1(k)$, is evident. In addition, I include

$$65536\left(1 - \left(1 - \frac{64}{65536}\right)^k\right)$$

as a continuous line. The match is relatively good, but the logarithmic scale hides the fact that the simulation falls detectably below the prediction for large k. This shows that the assumptions break down in this regime.

Having strengthened my confidence in the reasonableness of the approximations in the above analysis, I can now apply the results to the data from my LinkedIn experiment. According to the theory, Figure 1 and Figure 2 should both be small portions of a single graph representing the growth of a network with a maximum number of potential members N_{LI}.

$$N_{LI}\left(1-\left(1-\frac{F}{N_{LI}}\right)^k\right)$$

We determined F in section 2 to be about 78, all we need now is to perform a best fit of our data to find N_{LI}. The result, depicted in Figure 13, is that N_{LI} is surprisingly low, on the order of a few million at most. This is much lower than the advertised 25 million registered LinkedIn users.

There are several possible explanations for this discrepancy. First, my network may indeed be contained in a connected cluster of only a few million users. This is possible as I mainly invited or accepted invitations from people in one of my fields of interest, and did not pro-actively search for people far removed from these fields. Second, connectedness is defined only up to three hops, while fourth and higher degree connections are not considered in the LinkedIn framework. Third, the technique used to ascertain N_{LI} is not really able to determine the size of connected clusters. It estimates instead the size of the *strongly connected* cluster, neglecting isolated links that were not taken. In addition, the model is not being sensitive to links of my connections that I am not likely to use, further reducing the effective size.

8 Concluding remarks

Interestingly, my semi-empirical study proved all the tips given to me to be incorrect.

I was told that my network would increase ”exponentially”. Even were I able to keep up adding some number of friends per day, the network only grows linearly over time. Once I have most of my friends in my network, the network growth slows considerably to the organic growth rate which is sublinear in time. This result becomes even more reasonable when it is realized that the network size is already on the order of the size of it embedding connected cluster.

I was told that I would have a million connections in two weeks. It took about three times as much time (although others have told me that it took them much less). Even after a year I had not attained the two million mark.

I was told that I would find this network an extremely valuable tool. The value turns out to be only linear in the number of friends, albeit with a proportionality constant above 1. So while social networks are more valuable than simple contact list, they scale much more slowly in total network size than other types of networks. After a year of use, I find that I only use it once every week or two.

I was told that acquiring and maintaining such a network entails no cost. It turns out that the maintenance cost increases more rapidly with network size than the value, and that above about 500 friends maintaining the network becomes time consuming and cumbersome. After a year I find that I need to service requests (mostly invitations from people I turn down due to not meeting my criteria) at about the same frequency as I exploit LinkedIn for my own needs.

I was told that any two people in any of my fields of interest would be separated by no more than two or three degrees of separation. If tree-based models (natural for this kind of network) are to be believed, the degrees of separation between two randomly chosen people in my network are close to the maximum of 6, which incidentally is the value quoted as the separation between any two randomly chosen people. On the other hand, the network statistics I could gather does not match other models, such as random graphs or scale-free networks.

I was told that all 25 million LinkedIn users are connected (i.e., that there exists a single giant cluster to which essentially all users belong). In practice, my network seems to reside in a connected cluster of only a few million, from which I could not break out.

A Parsing LinkedIn Data

Computing the probability that two LinkedIn connections are themselves connected is not as easy as it seems. The idea is to retrieve and parse the connection lists of all friends and to look for duplicates. Retrieving web pages and extracting information from them is commonly called *web scraping*.

However, LinkedIn presents a challenge to web scrapers as it is based on AJAX (Asynchronous JavaScript and XML) programming. Viewing the source page of AJAX sites in the usual fashion reveals the Javascript source code, not the desired information. It is not difficult to retrieve the source of the HTML page as viewed, but the method to do so *is* browser dependent. Internet Explorer users viewing an AJAX generated page need only jump to the following address:

```
javascript:'<xmp>'%20+%20window.document.body.outerHTML+%20'</xmp>'
```

in order to dump the raw data to the screen for scraping.

The first thing one needs to know is that LinkedIn assigns a numeric key that uniquely identifies each user. You can readily see the key of any of your friends by jumping to that friends profile from your connection list. Look at the profile page's URL, it will be of the form:

```
http://www.linkedin.com/profile?...key=xxxxxxxx....
```

and that connection's key is readily seen. Finding you own key is a bit trickier. One way is to go to any of your pages such as your profile or connection list, dump the raw format as described above, and search for the string `key=`.

Now from the *my contacts* page:

```
http://www.linkedin.com/connections?trk=hb_side_connections
```

I can easily produce a list of all my contacts, their numeric keys, and the number of contacts they have. I can then form the URLs of the pages with their contacts, which are all of the form:

```
http://www.linkedin.com/profile?viewConns=&key=xxxxxx&split_page=n
```

with 60 contacts per page.

The program I used to automate this process is called *linkscraper*. The program starts with my contact list and extracts the keys of all of my friends. It also observes how many friends each friend has. After performing these steps, *linkscraper* creates the URLs of the contacts list for each friend, visits each of these links, dumps their raw format, and extracts the numeric keys of *their* connections (*my* second degree connections). The program then sorts the connections belonging to each friend creating a sorted list of numeric keys for each of my connections. It is now a simple matter of looping over every two different connections and checking if the two files had a numeric key in common. In addition, *linkscraper* calculates the number of connections in common between any two friends, building a triangular array that can be later analyzed in various ways.

References

[1] The American Mathematical Society *MathSciNet* Collaboration Distance calculator, online at http://www.ams.org/mathscinet/collaborationDistance.html.

[2] A.L. Barabási and R Albert, *Emergence of Scaling in Random Networks*, Science **286** 509-512 (1999).

[3] Tim Berners-Lee, *Giant Global Graph*, online at http://dig.csail.mit.edu/breadcrumbs/node/215.

[4] P. Erdös and A. Rényi, *On Random Graphs*, Publicationes Mathematicae **6**: 290 - 297 (1959); *The Evolution of Random Graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **5**: 17 - 61 (1960).

[5] The Erdös Number Project, online at http://www4.oakland.edu/enp/.

[6] Brad Fitzpatrick and David Recordon, *Thoughts on the Social Graph*, online at http://bradfitz.com/social-graph-problem/.

[7] Alex Iskold, *Social Graph: Concept and Issues*, online at http://www.readwriteweb.com/archives/social_graph_concepts_and_issues.php.

[8] J. Leskovec and Eric Horvitz, *Planetary-Scale Views on a Large Instant-Messaging Network*, online at http://www.cs.cmu.edu/ jure/pubs/.

[9] *About LinkedIn*, online at http://www.linkedin.com/static?key=company_info.

[10] B. Metcalfe, *Metcalfe's Law: A network becomes more valuable as it reaches more users*, Infoworld, Oct. 2, 1995.

[11] S. Milgram, *The small-world problem*, Psychology Today **1**: 61-67, 1967.

[12] A. M. Odlyzko and B. Tilly, *A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections*, online at http://www.dtc.umn.edu/ odlyzko/doc/networks.html.

[13] D. P. Reed, *Weapon of math destruction: A simple formula explains why the Internet is wreaking havoc on business models*, Context Magazine, Spring 1999.