Harnessing User Data to Improve Facebook Features

Greg Epstein
2010 Undergraduate Honors Thesis
Advised by Professor Sergio Alvarez
Computer Science Department, Boston College

May 12, 2010

Contents

Abstract The recent explosion of online social networking through sites like Twitter, MySpace, Facebook has millions of users spending hours a day sorting through information on their friends, coworkers and other contacts. These networks also house massive amounts of user activity information that is often used for advertising purposes but can be utilized for other activities as well. Facebook, now the most popular in terms of registered users, active users and page rank, has a sparse offering of built-in filtering and predictive tools such as "suggesting a friend" or the "Top News" feed filter. However these basic tools seem to underutilize the information that Facebook stores on all of its users. This paper explores how to better use available Facebook data to create more useful tools to assist users in sorting through their activities on Facebook.

1 Introduction

1.1 A Brief History of Social Networking

Since the advent of the commercial internet in the mid 1990s organizations and individuals have been exploring how best to utilize this amazing and unwieldy network for communication purposes. In the beginning email-like systems were developed to send messages between machines. Then came the creation of bulletin boards and newsgroups with the launch of systems like Usenet and Listserv. As the internet grew to resemble what it is today, website based approaches came to social interaction in the form of sites like Geocities and Tripod.com. These sites utilized the, then very popular, chat room paradigm to bring people together in real time chat. However none of these networks of communication resemble social networking sites as we know them today. In 2002 the site friendster.com launched as one of the first modern social networking services. This new type of communication system used the concept of profile pages to connect people to one another. Wildly popular, friendster gained over three million users in its first few months of operation. After proving the success of the concept, other companies quickly moved to launch their own services and by 2004 MySpace.com had come to dominate the market selling itself to NewsCorp in 2005 for $580,000,000. However after Facebook.com opened itself up to the general public in 2006, it quickly grew and overtook MySpace in 2008 with 132.1 million users a month. Of course much has changed in the social networking world since the early days of Friendster but Facebook still remains the largest social netwokring serve with over 400 million active users (see Figure 2 for details).

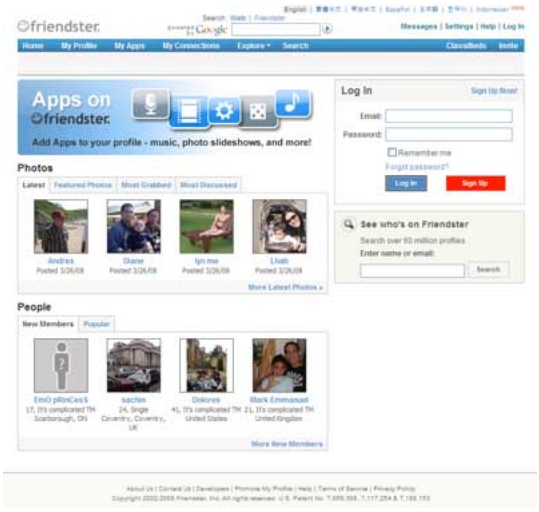

Figure 1: Original Friendster Frontpage

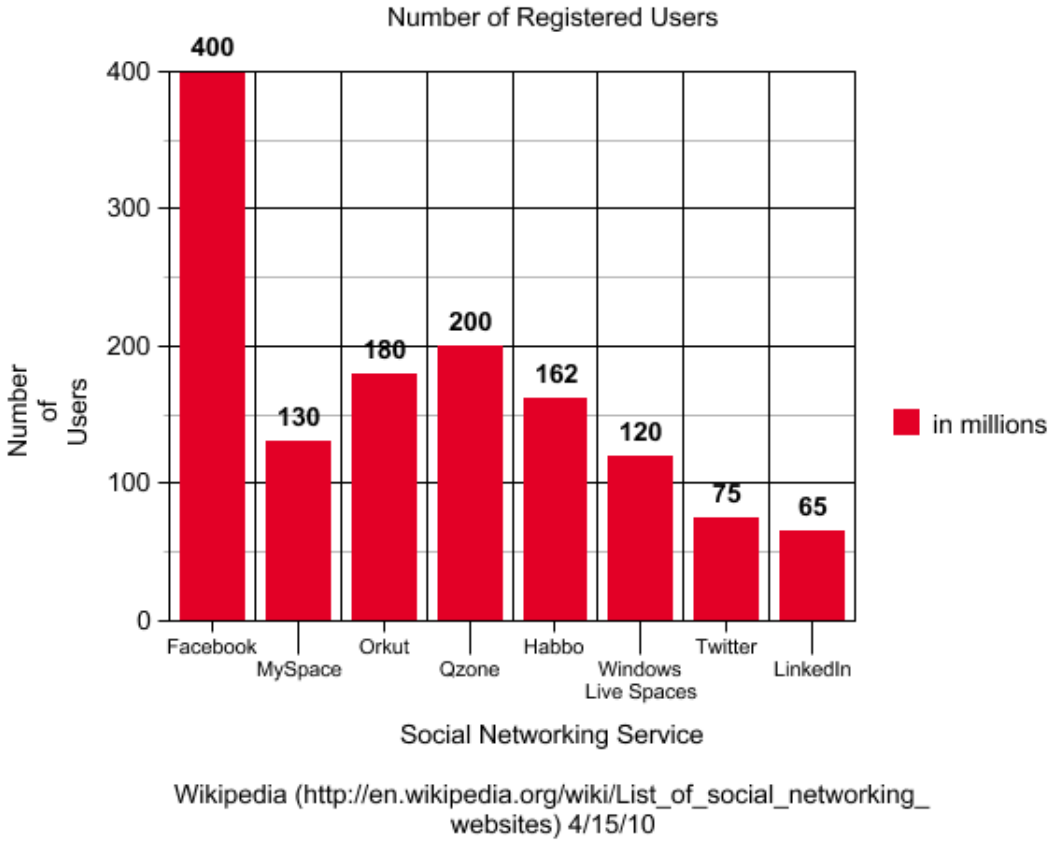

Figure 2: Number of Registered Users by Network

1.2 An Introduction to the modern Facebook

1.2.1 The Wall

For those that are unfamiliar, there are several major components that make up the modern Facebook site. Each registered user is given a profile page. The main component of each profile page is called "The Wall." A wall is a place where the user, or often time other users, can write things or leave links, photos or other media. The wall also automatically posts some of your Facebook activity. For example if you join a Facebook group or RSVP to Facebook event your wall will automatically post a blurb describing your activity. When someone posts something on your wall Facebook will notify you, or you can browse your friends walls to see what people have left there.

1.2.2 "Friending"

To gain access to someone's profile, for the most part, you have to send a "friend request" to that person asking for their permission to grant you access to their profile. Once they approve this request not only do you have access to their profile, but they have access to yours. This relationship is referred to as a friend relationship and has led to the use of the word friend as a verb to describe the action of submitting this request to someone.

1.2.3 Status Update

Another component of Facebook is the "Status Update." A status update is a piece of text, a photo, a link, or a few other rarer items that a user posts to update their friends on their own activity. Every status update is automatically posted to that persons wall. To clarify, posting a status update and posting something on your own wall are the same thing; however, users do not often consider their wall when posting status updates due to the way Facebook evolved and the current Facebook.com site interface.

1.2.4 News Feed

The third major piece of Facebook, for our purposes, is the "News Feed." This is one of Facebook's more recent additions and, while now an accepted and fundamental part of Facebook, was extremely controversial when it was first launched, and new research shows that users may still prefer the profile paradigm [10][11]. The news feed is an autoupdated list of your friends', and your, most recent activity. Everyones status update and wall activity is aggregated into your news feed so that instead of browsing peoples' profiles, looking for recent activity, you can just look at your news feed to see what people have been doing.

1.3 Facebook Suggestion and Filter Features

For whatever sociological reason, users of Facebook, especially at its inception, took great pride in the number of friends they had. To help users expand their number of friends, Facebook introduced a very basic feature that presents you with people that you might know as a list of suggested friends. While this feature does frequently show people that you know, its recommendation system is quite poor and is a feature that has drawn some criticism from more savvy users. It is this kind of recommendation system that we will explore in this paper by building a foundation from which these types of recommendations can be made. Another issue is because many Facebook users, especially in the 18-24 age group, have so many friends each of whom update their Facebook status multiple times a day, Facebook was forced to provide a feature on the news feed called "Top News." The top news feature is a tool that filters out your news feed to provide you with only the activity that it predicts you are interested in. Again, this feature never gained traction among Facebook users, in part because

Figure 3: Example Facebook Page

of its obvious low performance. The top news filter both filters out important content and lets through activities that users do not wish to see. This filter is another Facebook feature that we will improve upon by sorting through user data.

1.4 Facebook versus other Social Networking services

1.4.1 Bidirectional Connection

One aspect of Facebook that sets it apart from several other networking services, most notably Twitter, is its bidirectional linking of friends. This simply means that in the Facebook paradigm any connection between two friends necessarily goes both ways (bidirectional), while Twitter allows for users to "follow" other users without requiring the people they are following to follow them back. This distinction is important because it means a very different looking network graph between the two networks. In a Twitter-like system, a great amount of information can be extracted by looking at a users followers in relation to the people they follow, while the bidirectional set up of Facebook makes it harder to tease out this information. For example, it is easy to identify a celebrity in Twitter simply by noting the huge numerical difference in the number of people following them versus the number of the people they follow, while in Facebook it is harder to tell who is following who.

1.4.2 Privacy

Another major difference, research-wise, between Facebook and Twitter is on the level of privacy surrounding users activity. A vast amount of research has already been done with the Twitter data set as nearly all Tweets are available to the public. Conversely, Facebook has had to react to several privacy complaint episodes and has since allowed users to guard more of their information from the public as well as raise the default settings in the direction of greater user data privacy [12]. The result of these actions is that gaining access to large amounts of Facebook data has become more difficult and, I believe, has led many researchers to study Twitter despite the fact that, as mentioned earlier, Twitter is a fundamentally different network both in structure and in use.

1.5 Research Goals

There are two main specific goals of this project. The first is to create a friend ranking system that can rank a users friends from someone who they have great interest in to someone who may be a Facebook friend perhaps only for reasons of social pressure. The value a system like this would have is that it could be used as a base for Facebook's filter and suggestion features as well as well allow Facebook to perform more automated sorting algorithms to present a user data in a more consumable package. The second and related goal is to create an object ranking system and a corresponding object filter. These objects are the activities found on a user's news feed and the filter would be a substitute for

"Top News." Having an accurate understanding of Facebook objects could also potentially allow Facebook make their ads more targeted or allow users to sort through their friends' activity with greater ease. Remember that as Facebook use grows and more activity moves in the virtual realm it will become more and more difficult for users to manually sort through the increasing number Facebook objects and track their friends' activity.

2 Implementing Objectives

2.1 Friend Ranking

2.1.1 Mutual Friend System

Intro As earlier stated, one of the key elements in extracting information from Facebook's existing friend network data is creating a system to rank a users friends. The current Facebook paradigm provides absolutely no explicit tools for distinguishing between friends that you communicate with on a regular basis and that obscure acquaintance that you met one weekend several years ago. One attribute that Facebook does provide in its interface for every friend is the number of mutual friends you share with that friend. As the title suggests a "mutual friend" is a friend that both you and someone else are friends with. These shared connections are what gives someone's social network its shape and is the primary characteristic from which further data can be extracted. In its colloquial use, many users are interested in how many mutual friends they share with somebody as a barometer of how close they are to that individual.

The System With this use already in place the first friend ranking system that we implement is one where friends are listed in decreasing order by their number of mutuals friends.

Algorithm 2.1: MutualFriendRank(friend[]$friends$)

//Object type friend consists of a name and an int value representing
//the number of mutual friends
$sort(friends)$//sort by number of mutual friends
$reverse(friends)$//this puts the list in decreasing order
return ($friends$)

Friends that share few or no mutual friends end up on the bottom of the list, as they should as it is extremely rare that any of the users from which the test data was collected reported anybody "important" in this bottom echelon of the list. However the top of this list is less accurate, with only about 28% of users top10 friends filling the top 10 slots of their rankings, more on methodology and assessment can be found in the evaluation and analysis section 3.2 (The term top10 friends refers to a metric of evaluation where users marked their top 10 "preferred" Facebook friends)

Issues After these findings, interviews were conducted with the users of the tested data to explore why this algorithm failed, namely who was showing up on the top of the mutual friend list that shouldn't and why. The generalized findings in this area is that these friends at the top of the list that don't belong are mainly those that had so many friends themselves that by nature of having so many friends a large share of them just happen to overlap with the users own friends, not necessarily that both the user and the friend are close with many of the same people.

2.1.2 Mutual Friend Normalized for Popularity System

Intro The logical step that followed was to normalize for popularity, meaning to take into the consideration not only the number of mutual friends a friend has, but the number of total friends that friend has in order to weed out those friends that have a large number of mutual friends only because of their large number of initial friends. The first method implemented then is a ranking system not based on the number of mutual friends as in Algorithm 2.1, but instead on the fractional result from the following equation.

$$\text{normalizedValue} = \frac{(\text{number of mutual friends})}{(\text{number of friends})} \tag{1}$$

Improvements As it turns out however this system over compensates for the problem and eliminates any user with a decent amount of friends from the top of our rankings. In other words the number of friends overwhelms the number of mutual friends which is the value that holds the more relevant information. In order to place more weight back onto the number of mutual friends, since that is still the basis of this approach, the equation is modified to the form below.

$$\text{normalizedValueX} = \frac{(\text{number of mutual friends})^X}{(\text{number of friends})} \tag{2}$$

Adding the exponent X is intended to give the proper weight to the number of mutual friends and the result of this equation is what we call the "Mutual Friends Rank Normalized for Popularity." Giving X a value of two weighs the numerator too heavily, essentially leaving you with the same problem as the exclusive mutual friends approach in Algorithm 2.1. However after modulating the value of X and testing it on the provided data the best value for X turns out to be approximately 1.4. With this value the success rate rises from 28% to 39% (for more detailed results see section 3.2.4), a nontrivial amount to be sure, but still far short of anything extremely useful in terms of improving Facebooks filter and recommendation systems. The large problem with this approach is the wide range in the number of friends that people tend to have. Creating a unified system from this approach to deal with users with anywhere from 70 friends to 2200 friends, the rough lower and upper limit in the dataset, is near impossible because with such a range the ratio of mutual friends to total friends is not a sensitive enough indicator to accurately distinguish top10 friends from the rest.

2.1.3 Clustering system

Intro As the previous two approaches, described in section 2.1.1 and 2.1.2, reveal, using the mutual friends and total friends attributes for each friend on its own is not enough to create an accurate ranking system, luckily however there are more powerful tools that can examine the network as a whole and draw out the subsequent subtleties that the more narrowly focused methods described

earlier cannot. These types of methods involve creating meta characteristics about friends and groups of friends that can act as better heuristic than the explicit data that Facebook provides. The primary concept behind this approach is the idea of clustering groups of friends [15].

Harel Koren Algorithm The most useful of these algorithms is a multiscale layout algorithm which handles undirected graphs (remember that the symmetry of Facebook friends makes it an undirected graph), developed by David Harel and Yehuda Koren [1]. This algorithm essentially solves a clustering problem by separating a users Facebook friends into groups. The number of groups the algorithm produces is dependent on the nature of the network given to it. Each group represents a number of friends who are in turn mostly friends with each other. For example a group of High School friends would come out clustered together, while a group of coworkers could form its own cluster. Members of a cluster are mainly, but sometimes not only, friends with other members of the same cluster. While all this algorithm explicitly does is identify clusters we can use this new information to better tailor our friend ranking system.

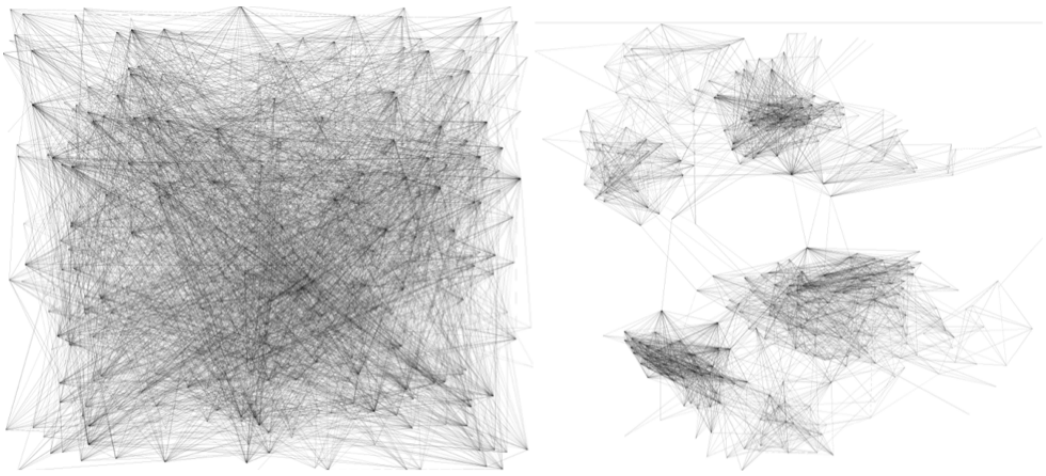

Figure 4: Left = Random Network, Right = H&K Clustered Network

Method The first step in this process is to identify the groupings that the H&K algorithm produces. Nodes that have one or few connections (i.e. friends with no mutual friends) are deemed not as their own cluster but as a group of nodes belonging to no cluster. Once K number of groups are determined, this value is often times between 4 and 10, they are sorted into an order. This order represents the relationship between groups. Groups that are adjacent to one another mean that they share the most cross group connections, i.e. many

members from one group are also friends with members of the other group. Note that the first and last group in the order are deemed to be adjacent.

> To clarify with an example, in an order of six groups as follows AFBECD, pairs AF FB BE...DA are most related while pairs AE FC BD...DB are least related. In terms of terminology pairs AF are said to have zero degrees of separation while pairs AE has two degrees of separation.

With these new attributes the equation in Equation 3 is developed.

$$\begin{aligned} &m = \text{mutual fiends normalized for popularity value} \\ &g = \text{number of cross group connections} \\ &d = \text{number of cross group connections with degree of separation} > 0 \\ &\text{Friends Value} = (m * 10) + g + d \end{aligned} \tag{3}$$

NOTE: Connections between groups with degrees of separation are counted twice, once as a regular connection g, and again as d in the last term.

Result While extremely simple, the Cluster Rank Equation above is amazingly powerful and surprisingly accurate. Note that it can only achieve this simplicity because of how powerful the H&K and ordering algorithms are in generating a meaningful heuristic. Upon closer inspection what the H&K algorithm really reveals is friends that share mutual friends across friend groups. The constant ten in Figure 3 is inserted to bring the mutual friends rank normalized for popularity value, which is almost always less than one, into the same scale as the other terms of the formula. With this more comprehensive approach the success rate in the top10 jumps to 71%.

2.2 Object Ranking

2.2.1 Intro

After achieving an acceptable friend ranking system we can now focus on the second objective which is to rank Facebook objects. These Facebook objects are the items that make up a Facebook news feed and include photos, links, status updates, wall posts, various application updates, and friending and group joining activities. The goal of this ranking system is to rank objects in such a way that a filter could be created in order to filter out all objects below a certain threshold. Currently Facebook employs such a filter called the "Top News" filter, however this filter has several gaping flaws which our own object filter will address. While the algorithm for Facebooks Top News filter is a proprietary information a basic amount of testing reveals its simplicity as well as its weakest areas. The three basic methods the Top News filter employs are as follows. First, objects with two or more likes or comments make it through the filter. Second, objects from

friends with a large number of mutual friends make it through, and third certain objects like links and photos make it through with less scrutiny than objects like status updates or wall posts. The first of these methods overwhelms the other two and accounts for what appears to be around 80% of items that appear on Top News filter. Of course, the problem with this approach, as well as the other two, is that it makes no use of the very useful friend data that we just showed one is able to extract from the network.

2.2.2 Point System

A common "mistake" that Facebook's Top News filter often makes is that it lets through items by a users friend that are heavily commented or liked by people that the user has no connection to. In response to this fatal flaw our ranking system will take into account the users associated with each object (either tagged, commented etc), apply points based on those characteristics, then find an appropriate threshold to allow objects with enough points to pass through the filter. Our method will also employ Facebook's principle, although to a lesser degree, of weighting pictures, photos, and links, heavier than text only objects since our research shows that users are generally more interested in those objects. Before going any further, the activity of "friending" is an exception to this entire system. If a user's friend befriends another of the user's friends, essentially adding a mutual friend between the user and first friend, the object of that activity will always pass through the filter. The following are the attributes an object can have and the points that that object receives for having that attribute.

- A friend likes the object: +.4 for every friend, +.9 for every top10 friend
- A friend comments on the object: +.8 for every unique friend comment, +1.7 for every unique top10 friend comment
  *in cases where a friend likes and comments an item only points for the comment are given.
- A friend is tagged in the object (either in a status update, photo or video): +1.2 if any friends are tagged, +1.7 for every tagged top10 friend
  *in cases where only a top10 friend is tagged 1.7, not 2.9, points are given
- An object has more than three comments by friends: +2
- A wallpost involving two friends: 1.8
- A wallpost involving at least one top10 friend: +2.1
- An object is a photo or video: +.9
- An object is a link: +.6
- An object originates from a top10 friend: +1.9
- An object has a comment: +.1 for every comment

- An object is liked: +.1 for every like

In our new system objects that have non-friends associated with them receive extremely few additional points. See figure 5

Figure 5: The above post receives points as displayed in figure 6

Reason	Points
Link	.6
Liked by Friend	.4
Commented by Friend	.8
Comments	.3
Likes	.1
TOTAL	2.2

Figure 6: Point Breakdown

2.2.3 Determining the Threshold

The next step is then to determine the threshold to which objects that exceed it can pass through the filter. Of course, here you are met with the classic problem of over and under blocking. If the threshold is too high, then content that the user wants to see gets blocked. Conversely if the threshold is set too low, then the user is bombarded with unwanted friend activity. After tinkering with the point system and examining the ROC plots the optimal threshold for

this scoring system is 2.4, with any object meeting or exceeding the threshold counted as a pass. Based on user feedback this new system improves the filter from a .475 success rate to .711 (according to our rating equation in section 3.4.2). The key to this improvement is both taking users feedback on what they want to see, and incorporating the findings from our top10 friend ranking system allowing objects associated with these friends to be distinguished from others.

2.3 A Filtering Example

To take an example of how our filter analyzes objects more accurately than Facebook's Top News feed, Figure 7 shows an object that passes the top news filter, that is marked by the user as unimportant, while Figure 8 is an object that does not pass the top news filter but should.

Figure 7: An object that passes the Top News filter that shouldn't

The object in Figure 7 receives .5 points as per the scoring system detailed in Section 2.2.2. The Top News filter lets this item through because it sees that there are five interactions associated with this object (three likes and two comments). However Facebook is unable to see that the poster of this comment is not a close friend to the user, or that this is one of dozens of posts that this friend made, or even that none of the commeters or "likers" of the post were friends of the user. Our filter however makes these important distinctions and gives only .1 for every non-friend user who interacts with the post. Figure 8, on the the other hand, does not make it through the Top News filter. It only has one outside interaction, the one user who "liked" it and Facebook therefore determines that it is not worth passing through the filter. Our filter however sees that both users involved in the wall post (the poster Greg Epstein and the postee Bruno Rodriguez) are friends of the primary user as well as the "liker" Andres Morales. Because these users are friends, as well as the fact that this post is of a video link, this post gets awarded 3.1 points, again as described in

Figure 8: An object that does not pass the Top News filter but should

Section 2.2.2. These two examples show just how important it is to parse the identity of the users associated with an object to determine the value of the object itself.

3 Evaluation and Analysis

3.1 Intro

The data used to build this entire project is comprised of 56 unique users. These users all volunteered their personal data by providing direct access to their Facebook account, the only way that Facebook provides for collecting the required information for this study. 28 of these users were University students or recent graduates, 10 were high school and middle school students while the remaining 18 were adults over 35. These proportions were chosen because they roughly match Facebook age group user data from November 2009 as provided by checkfacebook.com and shown in figure 9. This factor is important because

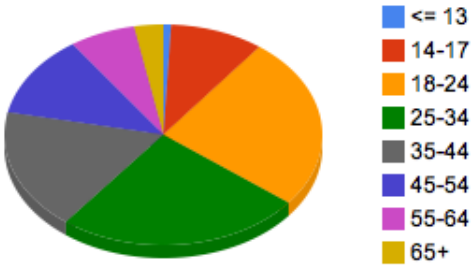

Figure 9: 2009 User Age Data from checkfacebook.com on April 1st 2010

not only does it help the data portray a more realistic landscape of users, but many of the older users have qualities in their accounts that were unique to their age group, namely that they had many times fewer friends. These unique factors made creating universal systems more difficult, but also made the final systems more accurate.

3.2 Friend Ranking Methodology

3.2.1 Intro

The first objective tackled was to create the friend ranking system. The end purpose of this objective was to create a filter to extract friends that the user cared to see the most about. For this reason we were less concerned about the evaluating the results at the bottom of the rankings and more concerned about the very top of that list. For this reason we developed the top10 friend metric. Like the name suggests the top10 friends is a list each user compiled that we would try to have the top 10 results of our ranking process mimic. Originally, when users created this list the friends on the list were ordered from one to ten; however when we began running tests we checked only to see if a friend was present in the list or not, ignoring what their ranking on the top10 list was.

This decision was made because the high amount of friends people have, make identifying a top10 friend equivalent to correctly selecting a friend in a pool that makes up approximately 2% of a users total friends, a sufficiently accurate metric making the rankings within that 2% make little relative difference.

3.2.2 Top10 Criteria

One major issue in giving the user the responsibility to create their top10 list was making clear to the user what types of individuals should be on this list. Of course simply telling the user to pick their "top 10" friends was greatly insufficient because users often chose friends or family members that they were especially close to, but who they did not interact with on a regular basis via Facebook [7]. Explaining the difference to users between a physical world friend network and the Facebook friend network is tantamount to receiving accurate top10 lists from participating users [4]. For this reason a set of guidelines was compiled and explained to every user before they created their top10 list. These guidelines are as follows

- top10 friend should be a friend that you are interested in following on Facebook
- top10 friend is not necessarily someone you personally trust or is close to you
- top10 friend is not necessarily someone you see everyday, in fact it may be someone you see rarely that thus communicate mainly through Facebook
- A top10 friend may or may not be someone who uses Facebook frequently

These guidelines at least gave us uniform feedback from the participating users.

3.2.3 Mutual Friend System

After these lists were gathered we first tested the mutual friend system. The system was run on the data, then the top 10 results that the ranking system produced was compared to the top10 friend list the user submitted, with the goal of maximizing the overlap between the two lists. The results from this initial comparison showed the following top10 results.

Average Top10 Overlap	Lowest	Median	Highest
28%	1	3	7

While this was the metric we used to measure success, in order to find ways to improve the system we looked at the data a little deeper. For each user we marked where their top10 friends fell within our ranking system. With this information we were able to tell if our ranking system was close to success (i.e. many friends listed in the top10 were in the top 50 rankings) or whether our system was filtering friends in a completely nonsensical way. As it turns out,

even from this crude initial approach of using only the mutual friend characteristic, 51 of our 56 users (91%) had their top10 present in the top 70 slots of our mutual friend ranking system. In other words if you were only interested in targeting the top10 friends, 91% of the time, you could cut the bottom 490 friends from the average user with 560 friends and still have a pool in which the top 10 friends were present (however we are interested in a broader consistent ranking system than just finding the top10). This finding indicated that in fact the mutual friend attribute was at least a valid starting point for proceeding our research.

3.2.4 Mutual Friend Normalized for Popularity System

After the previous data was collected interviews were conducted with a subset of the users to determine what characteristics existed in users in the upper echelon of our rankings that would allow us to filter them toward the bottom of the rankings. While very specific heuristics existed for individual users (i.e. friends over a certain age are never of interest), the only characteristic that became obvious across all users was that friends with a large number of friends gave them a better chance of having more mutual friends thus favoring them to an unfair degree. After discovering this phenomenon, we undertook a process to rectify this issue by adding the total number of friends a friend has into our equation creating the Mutual Friend Normalized for Popularity System. As mentioned in the Implementing Objectives section 2.1.2, this new method increased the success rate to the following.

Average Top10 Overlap	Lowest	Median	Highest
39%	2	4	8

This approach obviously improved the performance over the earlier method but also, because it was only a modular adjustment, seemed to show the upper limits to an approach based so heavily on the mutual friend metric.

3.3 Clustering System

3.3.1 Intro

In the process of conducting the user interviews regarding the Mutual Friend Normalized for Popularity System results, one very interesting tidbit emerged and became the basis for the clustering system. This small but extremely important observation was that top10 friends often had mutual friends from different parts of a user's life. Top10 friends frequently had mutual friends from a combination of several distinct groups, from high school friends to family to coworkers. While this information is interesting from a sociological perspective it is also invaluable for our purposes of friend ranking. To draw this information out of the data, however, was a bit more difficult.

3.3.2 Harel & Koren Algorithm

Everybody's Facebook network, as previously stated, is a bi-direction uniformly weighted graph [6]. What we needed to do was to group these nodes into clusters that existed naturally by the nature of the connections (i.e. friend relationships), but that were not immediately obvious. Several algorithms were tested before we discovered, and settled on, a clustering algorithm developed by Yehuda Koren, Liran Carmel, and David Harel [1]. This algorithm is an optimized clustering algorithm for large graphs. This algorithm finds vertices, in this case friends, that act as the vertex for each cluster that it determines then builds the remainder of the graph out from these verticies. Nodes that have edges to many of the same nodes are placed closer together while nodes with less connections in common get placed farther apart on the graph. While this algorithm supports 3d mapping, we only utilize to create 2d renderings. This algorithm works so well for our scenario because it draws out the subgroups in the network by grouping nodes that have a high number of connections between them. Using several extensions to Microsoft Excel, the program in which all data was collected, the Harel Koren algorithm produced a visual graph as illustrated in figure 10. This graph and much of this clustering process is implemented in NodeXL which both handles raw data and assists in illustrating the visual graphs [9].

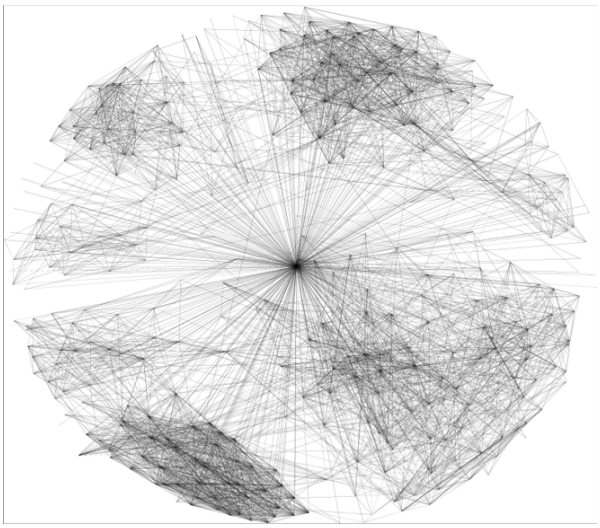

Figure 10: Graph Produced by H&K

The node in the center represents the user and has connections to every other node. Because we can already assume that all nodes displayed each represent a friend of the user's, we can artificially alter the data to remove the user from the network and produce a similar but more easily interpretable graph, showed

in figure 11.

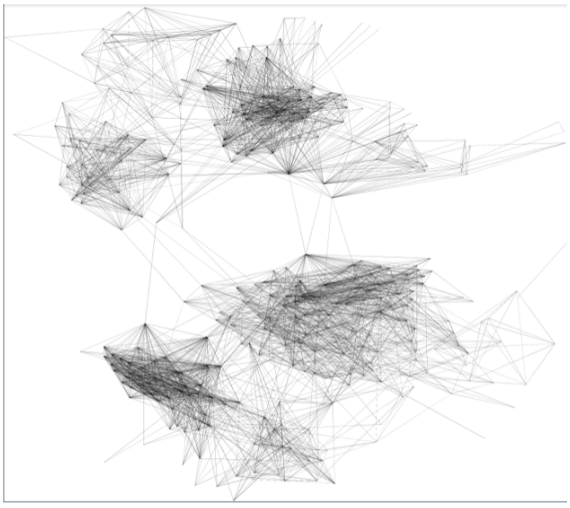

Figure 11: Altered H&K graph

At this point in the process some subjectivity is required to continue. The H&K algorithm, described in full detail in "A Fast Multi-Scale Method for Drawing Large Graphs" [1], produces an estimate of the number of clusters that exist in the graph along with which nodes fall into which cluster. The next step however is to annotate the graph in such a way to determine the following.

1 How many distinct groupings actually exist (this sometimes require merging to H&K clusters)

2 Which nodes fall into these groupings (this is determined by the H&K algorithm)

3 Which nodes have no associated group (this is done visually with some subjectivity)

4 What the appropriate associative order for the groupings is (associative order is described in greater detail in subsection Clustering System 2.1.3)

After annotating the graph, the grouping characteristics mentioned above are entered appropriately into our database as meta characteristics attached to each friend. With these new meta characteristics we compute the results from equation 3 and find the results are far and away better than the previous two methods. Using the top10 comparison metric our results jump to the following levels.

Average Top10 Overlap	Lowest	Median	Highest
71%	5	7	10

To further demonstrate the success of this method, we found that all but two of our participants had nine or more of their top10 present in the top 20 of our ranking when using this method.

3.4 Object Ranking

3.4.1 Intro

The Object ranking objective is to create a filter to apply to Facebook objects in place of Facebook's current "Top News" filter. These objects are what make up Facebook's news feed and include

- Status updates
- Wall Posts
- Photos
- Events
- Application Updates
- Shared Links
- Page Activity
- Friending Activity

3.4.2 Methodology

To evaluate the Object Ranking system each user evaluated their newsfeed once during a morning period and once during an evening. Each user would look at the 30 most recent items on the newsfeed and mark each object as either as an object they would like to see after a filter or as something they would rather have blocked. The number of objects users chose to have pass through a hypothetical filter ranged from 6 out of 30 to 20 out of 30. After collecting every submission, we had a total of 3360 objects on which to test our filter. Because the filter is supposed to make it easier for the user to browse material they care about, it is more appropriate in this situation to penalize overblocking of items greater than underblocking. For this reason the following metric was used to rate the success of the filter.

$$\text{Rating} = (\text{true positive rate}) - .75 * (\text{false positive rate}) \tag{4}$$

To maximize the rating value, we develop several point systems, create ROC curves and extract data from them to determine the best point system, then find the threshold that maximizes the above rating equation 4 [17].

3.4.3 Point System

In creating the the point system, the main tool of evaluation was examining true positive rates versus false positive rates of different filters via ROC curves. We first started by examining the accuracy of the Faceboook's own news filter results, which are shown in Figure 13

User #XX									
Object is a Photo or Video	0	0	1	0	0	0	0	1	...
Object is a Link	1	0	0	0	0	0	1	0	...
Friend Likes	0	0	2	0	2	2	3	3	...
Top10 Friend Likes	0	0	0	0	0	0	0	1	...
Regular Likes	1	0	1	4	0	1	0	1	...
Friend Comments	1	0	0	1	2	2	3	1	...
Top10 Friend Comments	0	0	0	0	1	0	1	0	...
Regular Comments	0	0	1	0	1	1	0	3	...
Friend Tagged	0	0	0	0	0	0	0	2	...
Top10 Friend Tagged	0	0	0	0	0	0	0	0	...
Wallpost Involving Two Friends	0	0	0	0	1	0	1	0	...
Wallpost Involving a Top10 Friend	0	0	0	0	1	0	0	0	...
Object from top10 friend	0	0	0	0	0	0	1	0	...

Figure 12: Example of Recorded User Object Data (each column represents one object)

	Actual Positive	Actual Negative
Filter Positive	1033	1073
Filter Negative	182	1072

Figure 13: Facebook News Filter Confusion Matrix

These results act as benchmark to any progress we hope to make. Combining these results into a more meaningful metric we extract the True Positive and False Positive Rates as shown in figure 14

	True Positive Rate	False Positive Rate
Value	.85	.5

Figure 14: Facebook News Filter Rates

With this information, along with informative statistics such as the ones in Section 2.2.1, we developed a point system that we thought mimicked the News Filter's general rules, and at the very least had a false positive rate of .5 when

the true positive rate was set to .85 (as shown in Figure 14) [5]. The ROC plot for this hypothetical News Filter is shown in Figure 15.

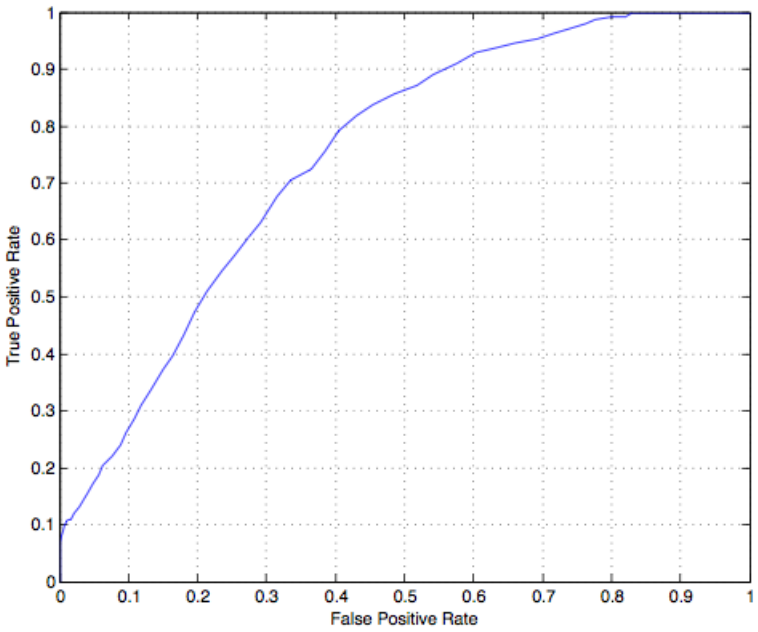

Figure 15: Hypothetic ROC of Facebook Top News Filter

After creating a point system that attempts to imitate the behaviors of Facebook's Top News filter, we move on to create a filter that can perform better. Finding statistics like the ones present in Section 2.2.1, but instead of being based on the Top News filter, based on the results from the users' feedback of an ideal filter, we develop a rough point system whose ROC curve, Confusion Matrix and Rates are shown in Figures 16, 17 and 18 respectively. While the process to determine the optimal threshold is described in the following section "Threshold", for the time being we calculate the various metrics holding the true positive rate at 85% to allow quick comparisons with Facebook's Top News filter which produces a true positive rate of 85% with our data.

From here, the point values are tweaked and replotted based on feedback from users' interviews and data observation until highest performing system, described in 2.2.2, is found, and shown in figure 19, 20 and 21.

3.4.4 Threshold

With our newly developed point system we next have to determine what the true positive rate that maximizes the rating equation in Equation 4 is. To do this we take the list of true positive rates and their corresponding false positive

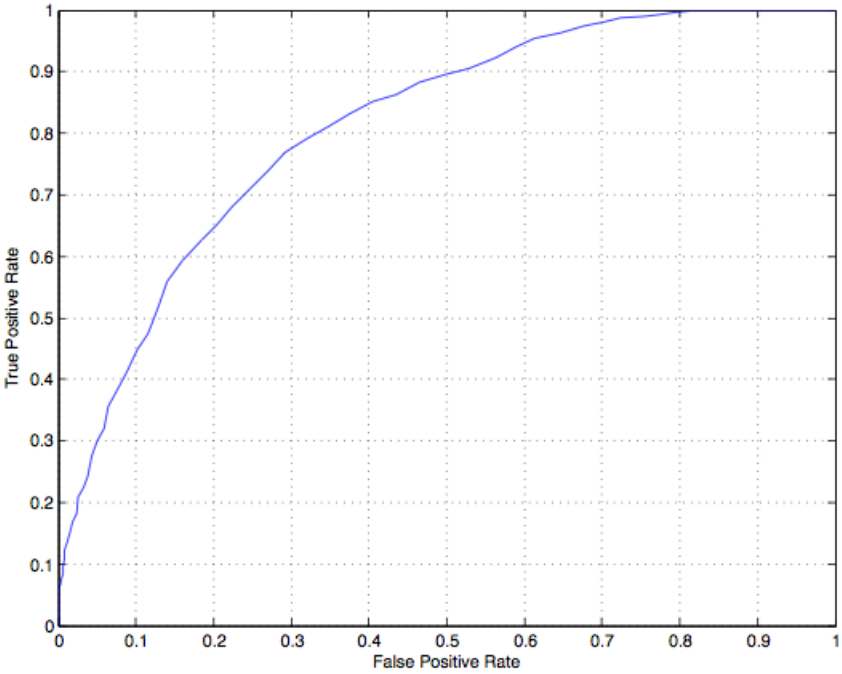

Figure 16: ROC curve of 1st filter attempt

	Actual Positive	Actual Negative
Filter Positive	1041	852
Filter Negative	174	1293

Figure 17: 1st Filter Attempt Confusion Matrix

rates that were used to create Figure 19, plug them into the above mentioned equation and find the maximum. An example of these calculations is found in the table in Figure 22.

Now that we know that the optimal true positive rate is 88.8%, false positive rate 26.2%, we have to find the corresponding threshold value to get these results. To do this all we do is take our 2145 object thats were marked by users as items they would like to have blocked, arrange them in decreasing order by their value determined by our new point system, then see that the value of the 562nd item (562 is 26% of 2145) has a value of 2.4.

	True Positive Rate	False Positive Rate
Value	.857	.397

Figure 18: 1st Filter Attempt Rates

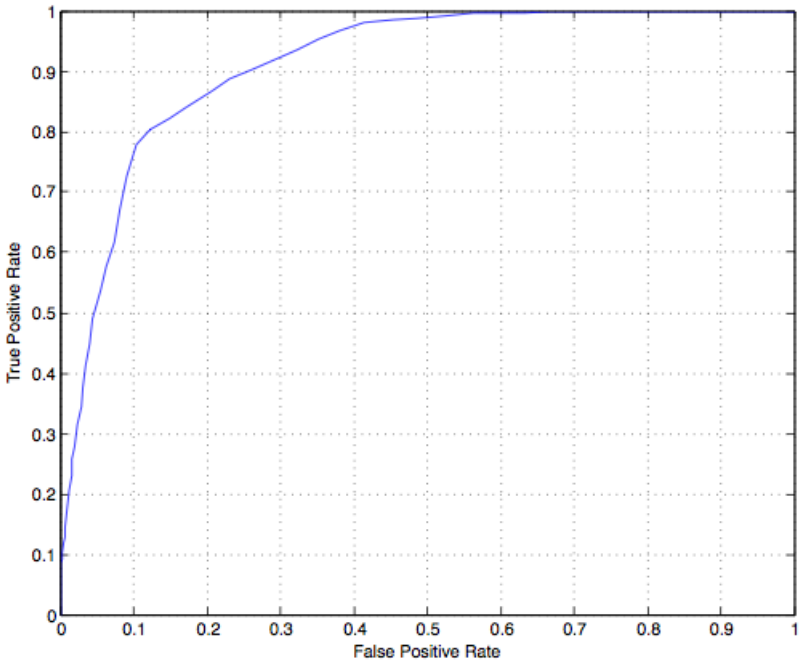

Figure 19: Final ROC Curve

3.4.5 Analysis

With our new point system in place it is important to compare the metrics of our new point system with those of Facebook's original Top News filter. To first compare the results visually, Figure 23 shows Facebook's Top News filter, our original rough point system and our final refined point system.

This data is shown as a histogram in Figure 24. Notice how the gap between the accuracy of the Top News filter and our final filter grows. To better illustrate this growing gap, Figure 25 graphs the difference between the false positive rate of the Top News filter and our final filter at any given true positive rate.

	Actual Positive	Actual Negative
Filter Positive	1041	852
Filter Negative	174	1293

Figure 20: Final Confusion Matrix

	True Positive Rate	False Positive Rate
Value	.866	.206

Figure 21: Final Rates

Line #	True Positive Rate	False Positive Rate	TPR - (.75 * FPR)
1	0.6181	0.0818	0.5623
2	0.6722	0.0901	0.6108
3	0.7262	0.1038	0.6586
4	0.7775	0.1221	0.6996
5	0.8031	0.1478	0.7115
6	0.8214	0.1709	0.7106
7	0.8397	0.206	0.7116
8	0.8663	0.2301	0.7118
9	0.8883	0.2621	0.7157
10	0.9029	0.2919	0.7064
11	0.9185	0.3218	0.6996
12	0.935	0.3517	0.6936
13	0.9542	0.381	0.6905

Figure 22: Maximize Rating Example. The maximum is found to be on line 9 meaning the ideal true positive rate is 88.8%

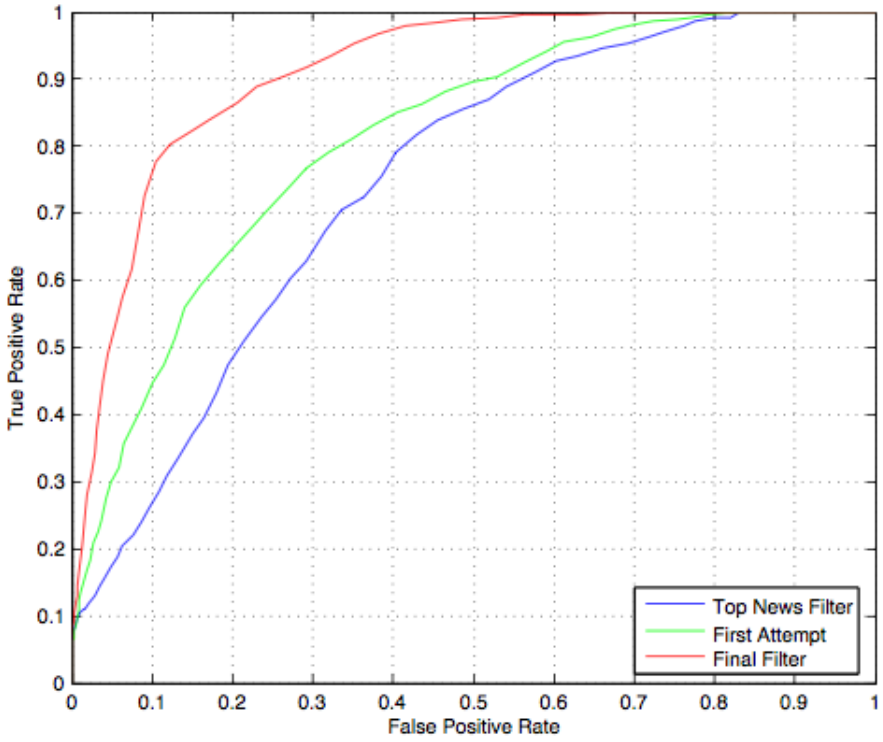

Figure 23: Comparing the ROC curves

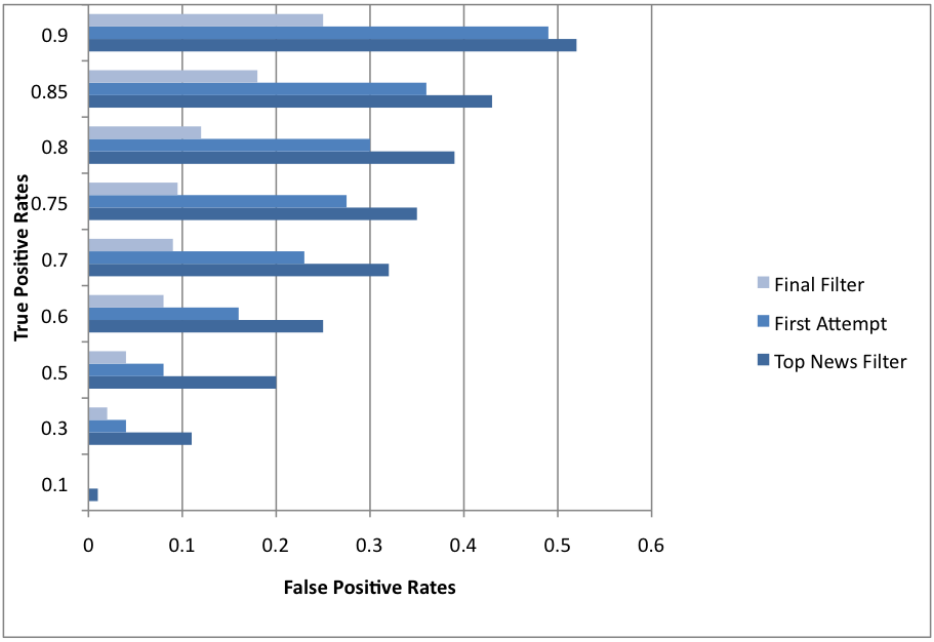

Figure 24: Atlernate Visualization for Comparing Filter Results

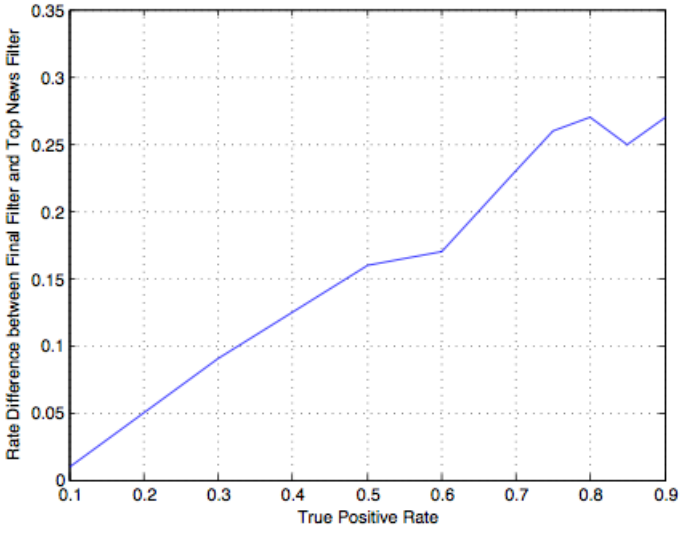

Figure 25: Gap between Final Filter and Top News Filter

4 Limitations

Our results above clearly show a leap in improvement on both friend and object filtering however, there are several limitations of our methodology and analysis that are worth discussing. The first and most glaring issue at the base of all our research is that we sample 56 unique users from a pool of over 400 million users. With a sample of only .000014% of the total user base it is obvious that our data is not fully representative [2]. The root of this problem is that unlike systems like Twitter, which originally had no private aspect to it as the idea was that you could post directly from your phone to the entire internet, Facebook has increasingly implemented layers of privacy walls around its users' information in response to growing concerns over data privacy [8]. These new privacy options, while both well intentioned and necessary for users, create greater barriers to data mining activities that could otherwise be done with the enormous dataset that the Facebook community provides [12]. While Facebook itself utilizes this mass of information in its ability to provide targeted ads to advertisers, it is very cautious in allowing even nonpersonal or unidentifiable information out to the public both for fear of a backlash from privacy advocates and for basic proprietary motives. However it should be noted that Facebook has indicated that it plans to lift many of these barriers in the next revamping of its system, a move many privacy advocates say may provoke legal actions in the federal court system by the FCC [18]. Another very basic element that is somewhat alarming is that our average user has 560 friends compared to the global Facebook average of 130. Despite the fact that the demographics age-wise between our test group and the global Facebook pool are the same, the average number of Facebook friends in our test group is 431% than what our average is expected to be. Another point of concern, in the process of friend ranking, is the subjective portion of the clustering system. Due to limitations in the programs used to implement the H&K algorithm, the final determination of how many clusters exist in a network is done solely on human judgement. While these choices are often easy to make, there are some occasions where the correct number is unclear without being able to process the data. This step may not impact the results to a large degree, but not being able to process the raw results of the H&K algorithm is a significant weakness in the process.

5 Conclusion

5.1 Friend Ranking

Friend ranking is an important base for a wide variety of Facebook recommendation and filtering systems and creating an accurate ranking system from the information that Facebook provides could impact not only Facebook's internal system but also the development of third party Facebook applications that could assist users in sorting through their friends activity. Both mutual friend systems taken in Section 2.1.1 and 2.1.2 perform surprisingly well for such a simple approach, but still short of anything to be useful in assisting users. Despite this, it is interesting to note that there is a basis for peoples' intuition that users with many mutual friends are more likely to be better friends of theirs, something that makes logical sense but is now substantively supported in this research. Another surprising result was that via the clustering approach we were able to predict on average 71% of users' top 10 friends. This number was much higher than initially expected, and with using so little user information, shows how accurate these type of systems could be if they took into account information that Facebook does not provide through its current API.

5.2 Object Filtering

With these positive results from the Friend Ranking system we integrated these methods into our object ranking system. Just from talking to Facebook users it is clear that Facebook's current Top News filter is unhelpful to say the least. The results from the initial object ranking system already outperformed Facebook's current filter leaving people to wonder why Facebook has allowed such an important feature to be so ineffective for so long (the Top News filter was introduced in the fall of 2009). Integrating the results from the friend ranking system we were able to decrease the false positive rate of the filter from 50% to 20% with a true positive rate of 85%. This 30% improvement signified a real accomplishment but optimizing our new filter we achieved a 26% false positive rate with a 88.8% true positive rate. The most telling visual of these results is shown in Figure 23. To gain further perspective keep in mind that 64% of objects should not pass the filter, so the fact that the false positive rate is only 26% is actually quite good.

6 Future Work

6.1 Friend Ranking

While the system we created was useful for the broad user base of Facebook, a much more effective ranking system could probably be created by incorporating the next level of user data. This data includes age, interests, network (aka university or geo location). Other research has already shown that the attributes that a user's friends offer up in their profile statement can often reveal key characteristics of the user [13] [14]. By utilizing these types datum it is probable that the friend ranking system could be enhanced to perform even better. To make the ideal friend ranking system, researchers with access to all Facebook's data could take into account user activity to track past interactions between users. With this historical information many of the predictive heuristics that other filters employ could be replaced with the actual records between users. This would allow for an amazingly dynamic system that could change the rankings of a user's friends on a monthly or even weekly basis as the user interacts with one group of people over another (i.e. perhaps a user interacts with a much different set of friends during the summer months when they are on vacation). While Facebook would never release these types of records to the public, it is something within Facebook's ability and, as the global network grows, may be something Facebook implements in order to track larger trends in user interaction.

6.2 Object Filtering

As mentioned above, this system is designed for all Facebook users and does not at all account for the great variety of Facebook users' preferences. Again, with Facebook's private records information they could create an unmatched filter. Either by implementing a simple "thumbs up thumbs down" feedback button for each object, where a user could express their preference for each item, or by using the existing "like" button on each object, Facebook could use this history to tailor the filter to individual users in a way that no other filter could. Similarly, Facebook could use the historical data to filter objects from users with frequent activity more strictly than those who only update their status on occasion. Another exciting way in which future object filtering could be implemented is through the utilization of Facebook's new "Graph API" [16]. This brand new API signifies a huge shift in Facebook's approach to integrating its services with the larger web. All information collected with the new features launched with the Graph API is permanently public and the explosion of the public Facebook dataset that is sure to follow will undoubtably change the landscape of how advertisers and researchers harness the power of tracking peoples' online activity [16].

References

[1] David Harel, Yehuda Koren, "A Fast Multi-Scale Method for Drawing Large Graphs", Journal of Graph Algorithms and Application, vol. 6, no. 3, pp. 179-202 (2002)

[2] Facebook Press Room Statistics, April 11th 2010, http://www.facebook.com/press/info.php?statistics

[3] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, 44-54. DOI= http://doi.acm.org/10.1145/1150402.1150412

[4] Ahn, Y., Han, S., Kwak, H., Moon, S., and Jeong, H. 2007. Analysis of topological characteristics of huge online social networking services. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 835-844. DOI= http://doi.acm.org/10.1145/1242572.1242685

[5] Cliff Lampe, Nicole Ellison, and Charles Steinfield, "A Familiar Face(book): Profile Elements as Signals in an Online Social Network", Dept. of Telecommunication, Information Studies, and Media Michigan State University, East Lansing, MI.

[6] John Breslin, Stefan Decker, "The Future of Social Networks on the Internet: The Need for Semantics", Digital Enterprise Research Institute, Galway, November/December 2007 (Vol. 11, No. 6) pp. 86-90

[7] Cliff Lampe, Nicole Ellison, Charles Steinfield, "A Face(book) in the Crowd: Social Searching vs. Social Browsing", Michigan State University, Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, Pages: 167 - 170, ISBN:1-59593-249-6

[8] Catherine Dwyer, Starr Roxanne Hiltz, Katia Passerini, "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace", Proceedings of the Thirteenth Americas Conference on Information Systems, Keystone, Colorado August 09 - 12 2007

[9] Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A. & Gleave, E. (2009), "Analyzing (Social Media) Networks with NodeXL", In C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies. Springer

[10] Casey Johnston, "Facebook users prefer profiles over newfangled(ish) newsfeed", Ars Technica, April 27th 2010 DOI=

http://arstechnica.com/science/news/2010/04/facebook-users-prefer-profiles-over-newfangled-newsfeed.ars

[11] Michael Arrington, "Facebook Users Revolt, Facebook Replies", TechCrunch, September 6th 2006 http://techcrunch.com/2006/09/06/facebook-users-revolt-facebook-replies/ (5/2/10)

[12] Brayden, "Facebook Research", orgtheory.net, Febuary 3rd 2009, http://orgtheory.wordpress.com/2009/02/23/facebook-research/ (5/2/10)

[13] Carolyn Y. Johnson, "Project Gaydar", Boston Globe September 20th 2009, http://www.boston.com/bostonglobe/ideas/articles/2009/09/20/project_gaydar_an_mit_experiment_raises_new_questions_about_online_privacy/ (5/2/10)

[14] Samuel Gosling, Sam Gaddis, Simine Vazire, "Personality Impressions Based on Facebook Profiles", University of Texas, 2007, ICWSM2007 Boulder, Colorado, USA

[15] Rafiq Phillips, Jacques Van Niekerk, Neb Kragic, "Social Graph" Facebook Application, http://apps.facebook.com/socgraph/ (5/2/10)

[16] Martin Gaston, "Facebook extends social network with Open Graph", April 23rd 2010, ZDNet UK

[17] T. Fawcett. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers", Hewlett-Packard Labs Technical Report HPL-2003-4, 2003.

[18] Nancy Weil, "The FCC's move, Facebook privacy issues redux", May 7th 2010, Businessweek