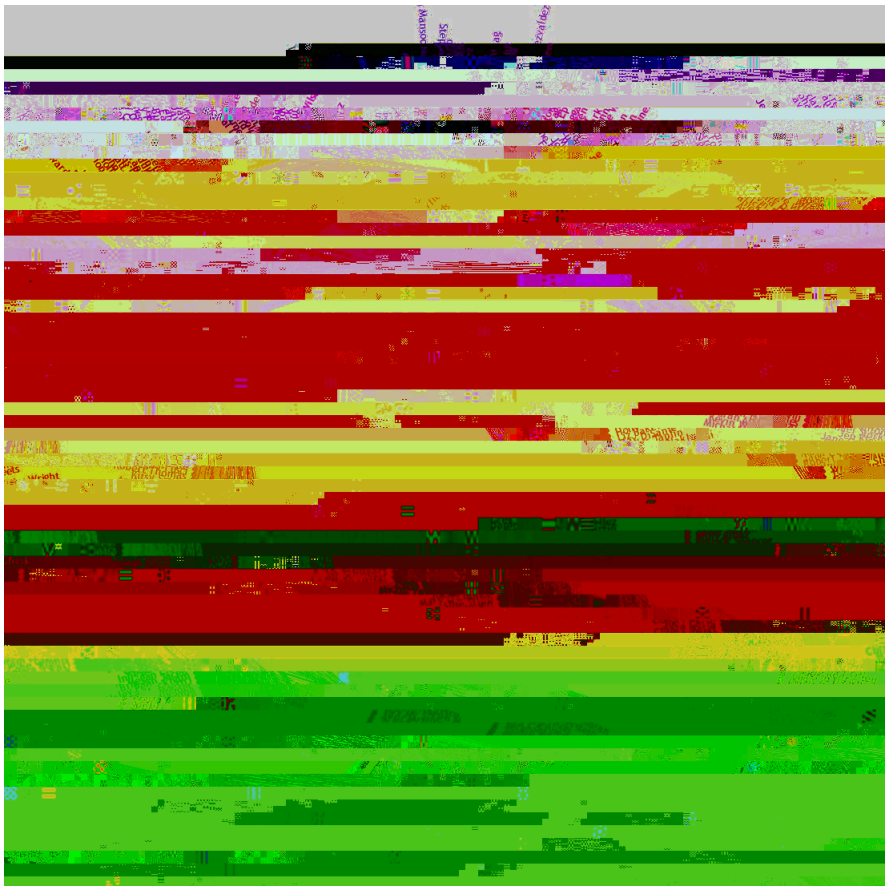
Harnessing User Data to Improve Facebook Features

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

May 12, 2010



Contents

1 Introduction

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

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.

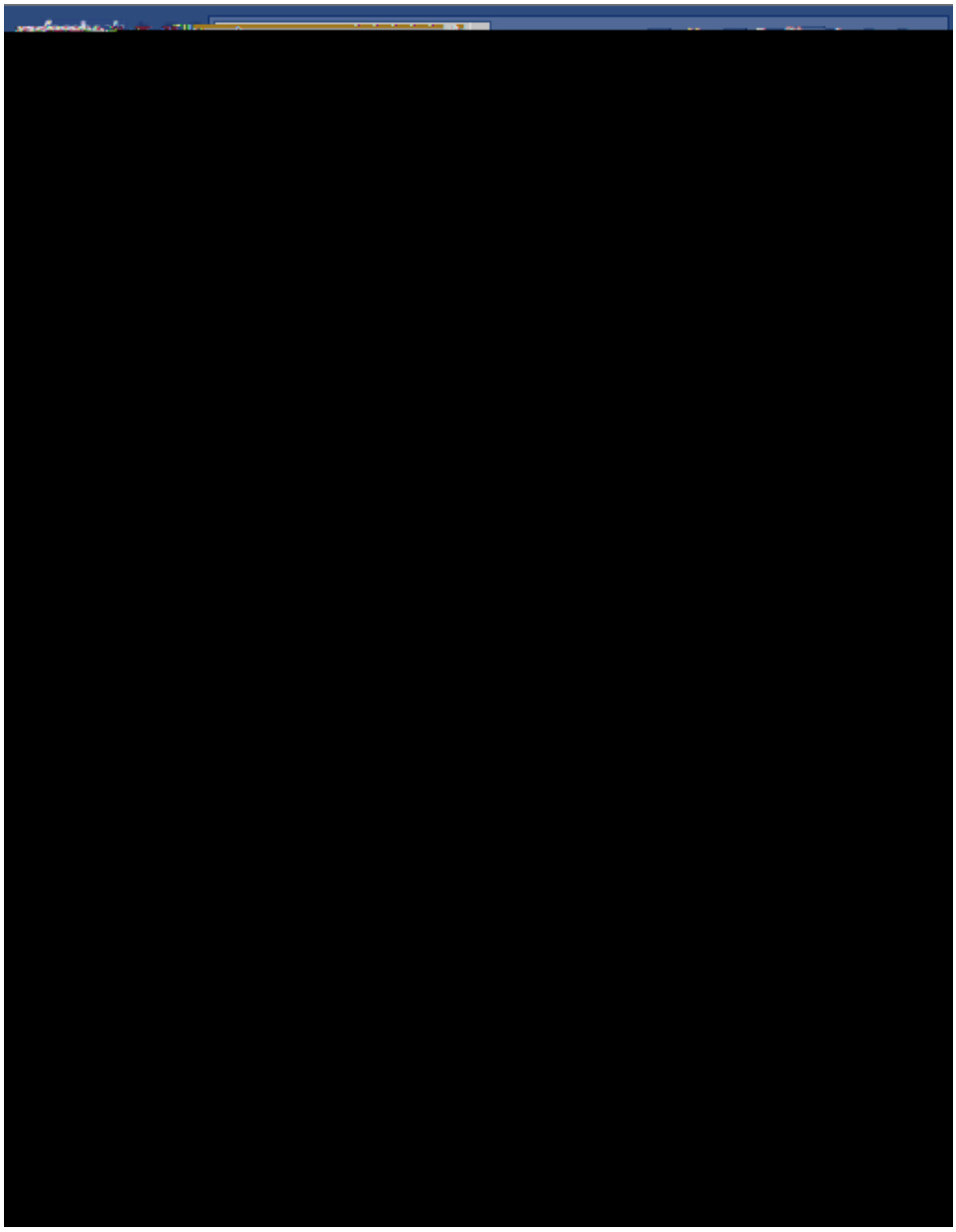


Figure 3: Example Facebook Page

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

"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.

earlier cannot. These types of methods involve creating meta characteristics about friends and groups of friends that can act as better heuristic than the

members from one group are also friends with members of the other group. Note

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

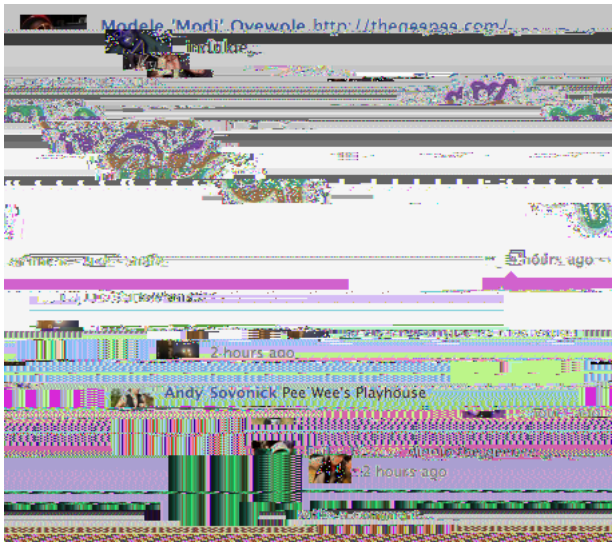


Figure 5: The above post receives points as displayed in figure 6

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

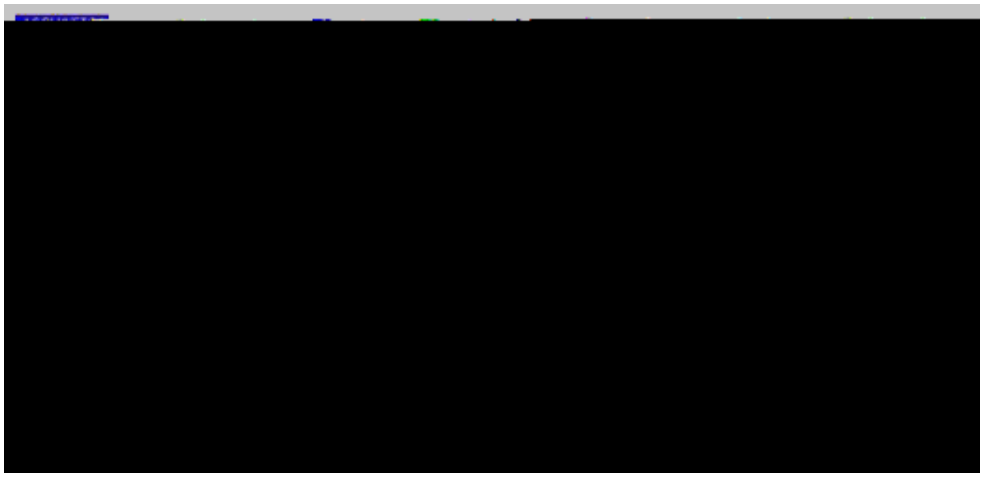


Figure 8: An object that does not pass the Top News filter but should

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.

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 su□ciently accurate metric making the rankings within that 2% make little relative di□erence.

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 insu□cient because users often chose friends or family members that they were
!

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.

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 annotat390.7(an)-0.81730.6Cetrot-334.4(ap)-0.4(t)-,8(4768)-0.8(t)-0.9(s)-0.439 groupin9chLarot-uCeerys

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.

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

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 di□erent 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.

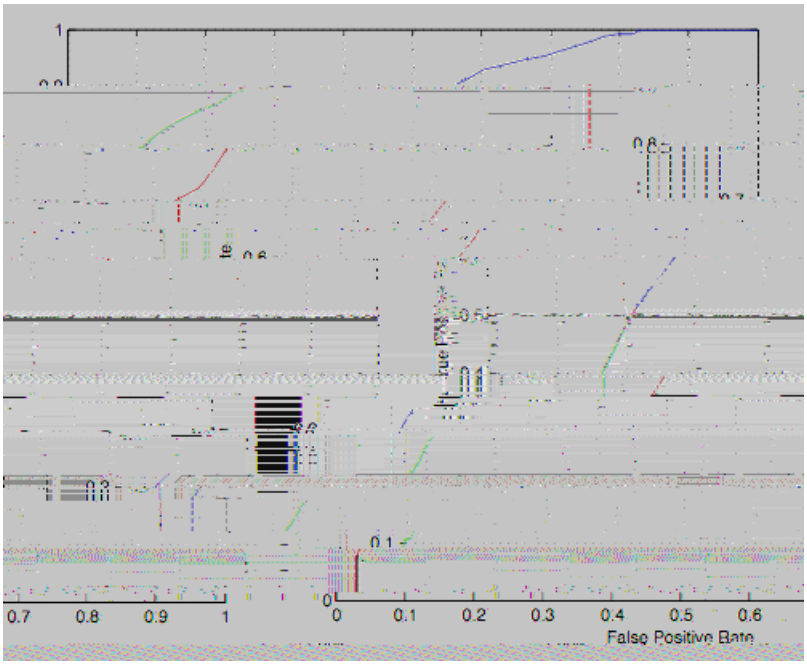


Figure 15: Hypothetic ROC of Facebook Top News Filter

After creating a point system that attempts to imitate the behaviors of

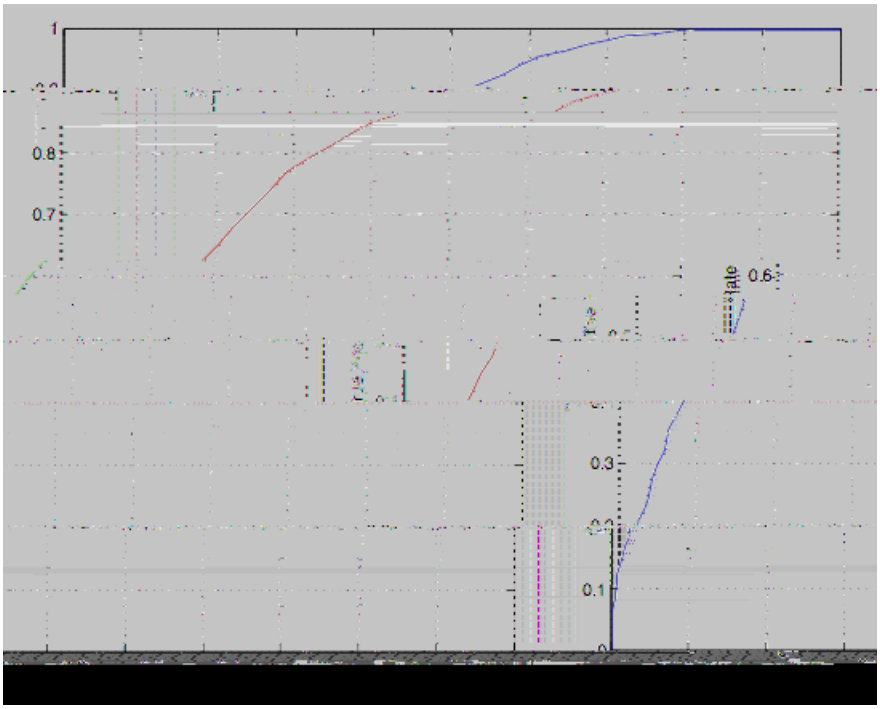


Figure 16: ROC curve of 1st filter attempt

	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 vale

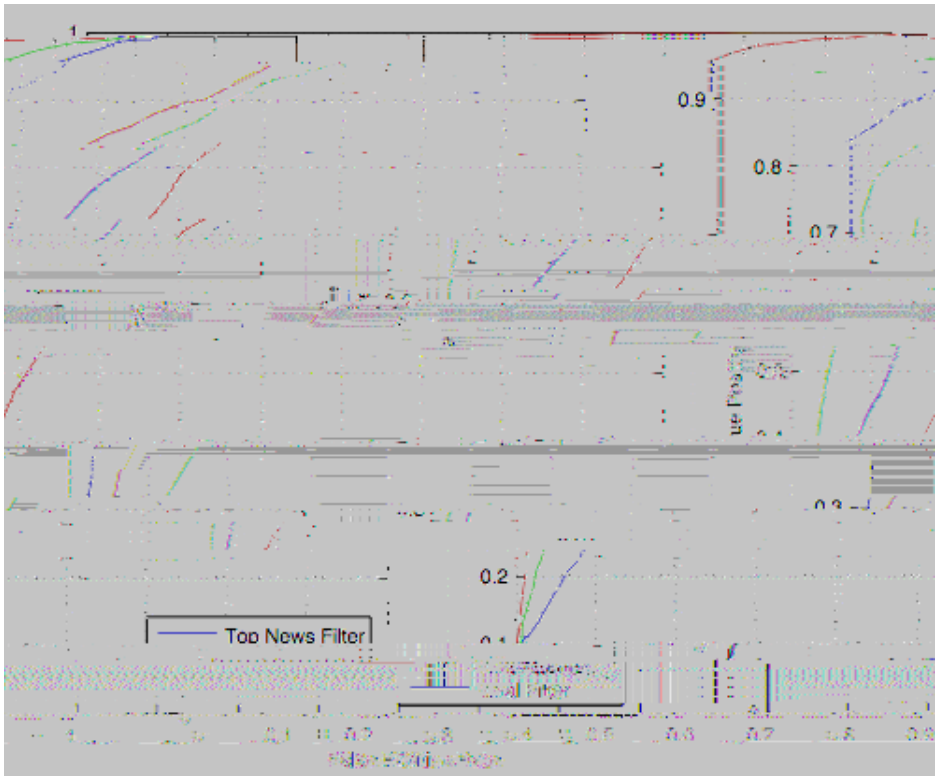
	Actual Positive	Actual Negative
Filter Positive	1041	852
Filter Negative	174	1293

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
0.9
0.8
0.7
0.2
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

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

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 (Ban□, Alberta, Canada, May 08 - 12, 2007). WWW ’07. ACM, New

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)