Quantifying Influence in Collaborative Social Networks

Elizabeth Bradley

erb2@princeton.edu

Advised by:

Thomas Funkhouser

Rework Sections:

**Set the Context. Explain the Main Goal of Research.**

1. Introduction (Motivation and Goal)
   1. Social Networks important, influence key to understanding
   2. Wikipedia is different, no prebaked social capital
   3. Attempt to understand influence

UGH. Introductions.

Abstract

*This paper attempts to analyze Wikipedia as a social network, and attempts to define influence and predict winners in administrator elections,*

# Great Example

With the emergence and rapid proliferation of social applications and media, such as instant messaging (e.g., IRC, AIM, MSN, Jabber, Skype), sharing sites (e.g., Flickr, Picassa, YouTube, Plaxo), blogs (e.g., Blogger, WordPress, LiveJournal), wikis (e.g., Wikipedia, PBWiki), microblogs (e.g., Twitter, Jaiku), social networks (e.g., MySpace, Facebook, Ning), collaboration networks (e.g., DBLP) to mention a few, there is little doubt that social influence is becoming a prevalent, complex and subtle force that governs the dynamics of all social networks. Therefore, there is a clear need for methods and techniques to analyze and quantify the social influences.

# Need to cite sources

Online social networks have dominated the online experience of the average user of the internet. – Networks Twitter filters our news, -- like Facebook determines our social connections. Key to understanding the power of the social networks is to understand those that have power in these networks. Identifying influencers.

This paper analyzes a social network not often covered in the literature: Wikipedia. Wikipedia is an online collaborative encyclopedia whose five million articles are the product of volunteer editors. There is a low cost of participation to edit, users do not even need to register an account.

The motivation is that Wikipedia needs more administrators to balance out things, develop a tool or understanding about what influence means to a Wikipedia user.

Wikipedia is different from other social networks in that it is collaborative and for the most part anonymous (although some will reference their education level or associated university). Unlike Twitter, individuals do not bring their social capital to the network, i.e. Lady Gaga is influential not because of her tweeting quality but because of her outside fame. Every editor is on the same theoretical playing field, which makes it a fascinating question as to what actions or user activity makes a user the most interesting

In this study, we examine factors of user history in relation to electability as admins, our definition of influence. We then attempt to predict the outcomes of RfA (request for adminships).

“[admins] have been peer-selected a s belonging to a class trusted with more power than a normal user.”

2. Related Work

A number of studies in have researched Wikipedia in attempt to understand it as a social network and specifically this admin / editor distinction:

* Power of the Few vs. Wisdom of the Crowds – parses out influence of the administrator class.

Other studies have attempted to quantify influence (not qualify)

* Everyone’s an Influencer: Quantifying Influence on Twitter, uses word-of-mouth diffusion to quantify influence. Basically “who listens to whom.”

3. Data and Implementation

4. Quantifying Influence

This section discusses the impact of different categories on calculating influence. The machine learning algorithms

We do have a ground truth measure, so we can accurately assess the effectiveness of prediction methods.

PCA tells us how the users are different because we can assume that those running for election all have very similar profiles, we want to see the areas where users are the most different. PERFORM PCA ANALYSIS HERE, use it to order the areas (4.1,4.2,4.3,4.4).

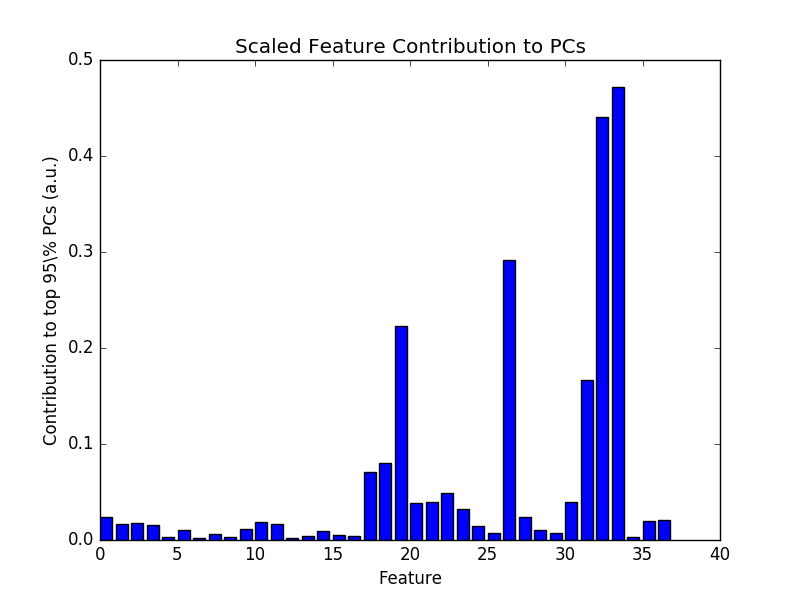


Figure 4.0.1: Scaled Feature Contributions to PCA

The two largest values are for featured content (primarily because most users have no featured content

4.1 Activity

TXT:'''Support''' User has been around since Nov 2004 and had the first RFA in 2006 and has waited for 2 years to try again after overcoming the points raised in previous RFA. User has more than 19000 mainspace edits and over 25000 overall.

Figure 4.1.1 Excerpt from Pharaoh of the Wizards vote for Ynhockey (2008)

There are a few sledgehammer measures that are often relied upon that calculate the general user activity (rate of activity and quantity of activity). These measures include number of pages created, total edits, days active from first edit to the date of election, number of accounts created (different languages and different areas of the site), votes in previous RfAs. Many of these measures are quite large, but can be affected by bots (which is why they aren’t a be all end all).

Graph, Max, Mean, Std. Dev.

Figure 4.1.2 Visualization of Edits

However, much of this occurs because of edit bots (which is looked poorly upon), so we remove outliers (with a factor of 5.0 on interquartile range). RUN ALGORITHMS ON BOTH, a tree algorithm , Bayesian, and ground truth rule.

Perform Feature importance for which is the most important feature

4.2 Role on Site

TXT:'''Neutral'''. I'd support in a heartbeat if you had two or three times as much experience in the Wikipedia namespace.

Figure 4.2.1 Excerpt from Useight’s vote for AlbertHerring (2008)

Justification.

Wikipedians love discussing roles (Wikipedia dragon, etc.), fitting that some roles would correspond with being “admin-like.” Article (aka mainspace edits), user talk page edits (editor-to-editor chats) , Wikipedia edits (meta things like voting etc.), category, template, etc.

Graph Max, Mean, Std. Dev.

Figure 4.2.2 Visualization of Edits

Analyze w/ outliers

Analyze w/o outliers

Perform feature importance to explain which namespaces are the most important

4.3 Distinction (good articles / lists)

Pull-Quote

TXT:'''S'''upport. Graham is a content beast. I think all content beasts, especially in the field of GA/FA content, should have the option of admin tools provided they are unlikely to abuse them. This content beast can be trusted.

Figure 4.3.1: Excerpt from Jfdwolff’s posisive vote for GrahamColm (2008)

Justification

Wikipedia has a measure of article quality. Most pages are owned by “projects” which rate and maintain pages. Pages are ranked stub, start, C, B, A, and then Good. No record exists for Good Article promotion (since it occurs project-by-project, most of which do not keep records) except self-reporting. Therefore, I looked to another measure of distinction: Featured Articles. On the Wikipedia homepage (<https://en.wikipedia.org/wiki/Main_Page>), one article appears prominently as the featured article of the day. Of the five million articles on English Wikipedia, nearly 4,700 have been featured, thus approximately ~0.1% are of high enough quality. (<https://en.wikipedia.org/wiki/Wikipedia:Featured_articles>).

Although not required, all featured articles are Good and the user who puts them forward for consideration is a key author. Therefore, a user “with” featured articles and lists may be considered a content creator, and a very good one at that.

Quantitatives

Only predicts 53%, probably add on to others.

4.4 Behavior (AfD votes, votes positive or negatively)

TXT:'''Support.''' An experienced and level-headed editor with a solid mainspace and projectspace contribution record, including substantial AfD participation.

Figure 4.4.1 Except from ---‘s vote for --- (2008)

Do people like hardliners?

4.5 Location (Break up edit counts, centrality measure)

TXT:'''Oppose'''. Sorry, but I'm going to have to oppose based on lack of mainspace contributions. Now, I say "contributions" instead of "edits" because you have 17,000 edits there. However, if you take a look at [[User:Useight/RFA Standards|my admin criteria]], I'm not impressed by endless clicking using an automated script. Huggle has done wonders for your edit count, but it's going to take communication with other editors, work in the Wikipedia namespace, article contributions (on the article you've edited most, you've only edited it 14 times). Communication is a very important attribute for an admin to have, as others have mentioned. I'm a very big supporter of mainspace work, so I'll have to recommend you do some actual content-building work there, as well.

Figure 4.5.1 Excerpt from Useight’s vote for Altias

5. Conclusion

Write when answers have been obtained.

6. Future Work and Expansions

Further analysis could be done that reads activity with a more fine-toothed comb, perhaps with sentiment analysis as to whether the user…

7. Acknowledgements

Thomas Funkhouser

Cloud9

8. Honor Code

I pledge my Honor that I have not violated the Honor Code during the writing of this paper. /s/ Elizabeth Bradley

9. Citations

Figures Wanted:

1. To show the influence of years, “better” candidates over time? Argument of time as a force?

X: Time of Election

Y: # Edits

Color: Successful v. Non-Successful

2. Regression Tree / Random Forest / Bagging virtual tree. (Like Word of Mouth.pdf pg. 6)

3. Charts of Mean / Median / Mode, etc. for each of the activity types.

4. Bar Chart of Prediction with each type (something, bagging, randomforest) for each section with a line charting its decreasing error over time.

Appendix: Glossary of Terms

Featured (Article / List)

Article

List

Editor

Administrator

Articles for Deletion (AfD)

Namespace

Mainspace

Wikipedia

MediaWiki

Category

Order of Section 4:

Because of PCA, Featured Content – Implementation, Performance, Analysis

The following is a possible outline for your paper.

\subsection{Introduction}

\begin{itemize}

\item Problem statement

\item Motivation and goal...The goal of this project is...

\item Roadmap: The remainder of this paper is organized as follows....

\end{itemize}

\subsection{Related Work}

\begin{itemize}

\item Survey of prior work with similar goals

\item Comparison to your project

\end{itemize}

\subsection{Approach}

\begin{itemize}

\item Key novel idea

\end{itemize}

\subsection{Implementation}

\begin{itemize}

\item Things you implemented. How you did it?

\end{itemize}

\subsection{Evaluation}

\begin{itemize}

\item Experiment design...

\item Data...

\item Metrics...

\item Comparisons...

\item Qualitative results...

\item Quantitative results...

\item Further results needed...

\end{itemize}

\subsection{Summary}

\begin{itemize}

\item Conclusions...

\item Limitations...

\item Future work...

\end{itemize}