Finding Trendsetters on Yelp Dataset

Pierfrancesco Cervellini[1]
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada
Email: pcervell@lakeheadu.ca

Angelo Garangau Menezes[1]
Department of Mechatronics Engineering
Tiradentes University
Aracaju, SE, Brazil
Email: agaranga@lakeheadu.ca

Vijay Kumar Mago
Deparment of Computer Science
Lakehead University
Thunder Bay, ON, Canada
Email: vmago@lakeheadu.ca

***Abstract*—The search for *Trendsetters* in social networks turned to be a complex research topic that has gained much attention. The work here presented uses big data analytics to find who better spreads the word in a social network and is innovative in their choices. The analysis on the Yelp platform can be divided in three parts: first, we justify the use of Tips frequency as a variable to profile business popularity. Second we analyze Tips frequency to select businesses that fit a growing popularity profile. And third we graph mine the sociographs generated by the users that interacted with each selected business. Top nodes are ranked by using Indegree, Eigenvector centrality, Pagerank and a Trendsetter algorithms, and we compare the relative performance of each algorithm. Our findings indicate that the Trendsetter ranking algorithm is the most performant at finding nodes that best reflect the Trendsetter properties.**

***Index Terms*—Yelp, Early Adopters, Trend Setters, Big Data, Social Network**

1. Introduction

After 2015, the number of business reviews on Yelp reached quota 90 million reviews [1]. As a crowd sourced review system, Yelp offers its users the ability to express their opinions on local businesses, and build social networks; users “friend” each other, build directed graphs based on apparent shared preferences thus generating information about the influence patterns they present. Who found that business first? Who spread the word about it the most? The answer to these queries is of great relevance to businesses: it allows them to maximize the return on their marketing budgets and identify the best target audience for their social media strategy. And within a social network, *Trendsetters* (TS) act as multipliers of information distribution rates and as a result make for an ideal audience.

In this paper we took on the 2015 Yelp Challenge dataset in search for TS using the social graph analysis algorithm by Saez-Trumper et al. [4]. Following their work, we define TS as innovators, people who have a tendency to pick up on “the next big thing” before it becomes popular, who have the ability to propagate information quickly.

To identify TS in the context of Yelp, potential businesses are identified by a positive change in their popularity curve, and then the slope of the curve over a period of time is tracked. The selected businesses exhibit the sharpest positive slope change. Next, a business is represented as a collection of trending expressions (such as the business name), and together with the time-stamp of when the users made such references we compute a weighed innovation graph.

Finally, a TS ranking is computed and compared with the ranking resulting from Pagerank algorithm, and graph centrality measures. A large body of work exists on this topic but in most cases the research focuses either on graph topology to rank users based on node properties, or temporal information diffusion. Other than the paper by Saez-Trumper et.al, no study seems to have combined the two approaches. Saez-Trumper et.al analysed Twitter and built the social graph using hash-tags as their screening tool for inclusion/exclusion. Whereas their conclusions relate to a virtual kind of trendSetting, one with an very low level of commitment, the proposed analysis tries to infer a pattern in a type of influence strong enough to convert an opinion into business revenue by driving people through the doors of an establishment.

The principal contributions presented in this work can be summarized as follow:

- The analysis of a well known social network (Yelp) utilizing a TS algorithm to yield its inward features.
- A comparison of the most used ranking algorithms in order to show their efficacy in a quantitatively and qualitatively approach.

The rest of the paper is organized as follows: Section 2 presents the related works, followed by the dataset definition in Section 3. The experiments are presented in Section 4 and the results in Section 5. Finally, conclusions and discussions are presented in Section 6.

2. Related work

2.1. Local Experts

There is already some published work based on the Yelp Dataset challenge where the author uses a set of features

[1] *Pierfrancesco Cervellini and Angelo Menezes are co-first authors*

to determine whether a business is popular in an specific area [3] and whether a person is a “local expert” or not [6] which can be used for generating weighted reviews for businesses and improvement of recommendation systems. This shows that an ecosystem of users exists on the platform, and that among them, an informal structure based on perceived expertise and credibility affects the spread of information.

2.2. Early Adopters

Looking at data from a popular on-line virtual world (Second Life), E. Bakshy et al. [5] discovered that TS are usually users with few friends that generally play less hours than the average. This is an important observation because it shows that users that could be considered outsiders in a trivial analysis, gain importance when the time factor (to be an early adopter) is considered. In our work with businesses on Yelp, we follow the analytic flow proposed by Saez-Trumper et al. [4] to differentiate those that create cascade behavior.

2.3. Network topology based algorithms

The idea of using a Pagerank algorithm to analyze qualitatively how an user is influential in Social Media was developed previously by Weng et al [7]. M.Cha et al. [9] evaluate users’ influence on Twitter, based on node in-degree, retweets, and mentions. Their main conclusion is that while the number of followers (node in-degree) is a measure of popularity, it is not necessarily an indicator of influence. They called this: “The Million Follower Fallacy”. Even though this study considers the time dynamics of communication, it only examines them as they relate to the staying power of a user’s influence over the network, and do not include a screening by topic of interest.

2.4. Temporal Factor

In 2008, A. Anagnostopoulos et al. [10] published a study that looked at three proposed reasons for congruence between actions among users in a social network: environmental factors, homophily [8] and influence. Environmental factors refer to the ability to post about a specific event that a group of users has access to because of geographic location; homophily refers to users posting about a topic because the topic itself is a common interest. And with regards to influence, they conclude that it can only really be discussed in the context of a social network if a time causality is paired to the actions among users [10].

Another major limitation of looking at a static network topology is that it does not describe the propagation speed across edges which may be faster or slower. A case in point being information that can travel slower across a single edge than it can by traveling through multi-node paths that use faster edges [11]. While these studies offer great insight, unfortunately they do not propose a ranking function or a way to include only nodes that have shown interactions with a specified topic.

2.5. Our Approach

Our analysis seeks to compensate for this very issue by using an algorithm that combines temporal and topology analysis. Both the algorithm and our definition of TS stem from the work of Saez-Trumper et al. We derive our directed social graph by including users that have shared content related to a specific topic, but we apply this analysis in the Yelp context (Yelp businesses in this case).

3. The Dataset

3.1. Data Source

The dataset is provided by Yelp as part of the 5th Yelp Dataset Challenge. It consists of 1.6M Reviews, and 500K Tips by 366K Users for 61K Businesses. Businesses include 481K attributes (hours, parking availability, ambient), and there are a total of 2.9M social edges, as well as aggregated check-ins over time for each of the 61K businesses [2]. The dataset is freely available at http://www.yelp.ca/dataset_challenge.

3.2. Segmentation

To facilitate computation we opted to build and test our model of analysis on a 10k randomly picked large sample of business listed in the Yelp dataset. Using the partial dataset, all the chosen businesses were analyzed and segmented according to their features in order to observe which of them had reached our considerations.

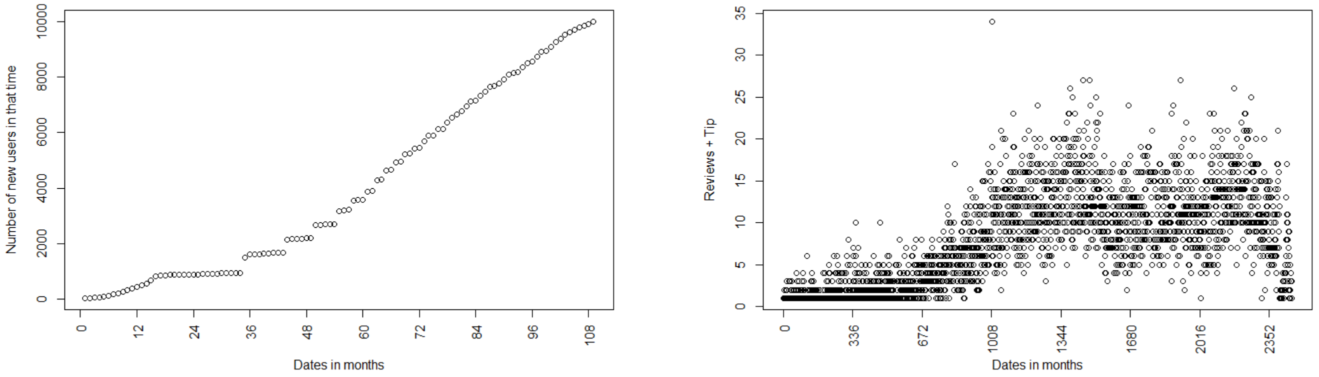
4. Experiments

4.1. Measuring Popularity

To capture data reflecting modern Yelp user behaviour we plot the monthly count of Reviews and Tips. Using Figures 1a and 1b as guidance, we chose 2008-11-30 as a historical cut off point since it yields the consistent growth of Yelp in relation to past years. We also exclude businesses with less than 20 reviews, and we perform outlier detection on aggregate Reviews, Tips, and Checkins with the use of box plots as shown in Figure 2 for the better establishment of a model.

Initially we made the assumption that just the number of Checkins a business receives can be used to measure its popularity. However, Checkins do not have a full timestamp and as a result we had to find a different variable to measure popularity over time. To help identify one, we use the Pearson Correlation coefficient, which shows a measure of the strength of the linear relationship between two variables, in order to compare Reviews, Tips, and Checkins as shown in Table 1.

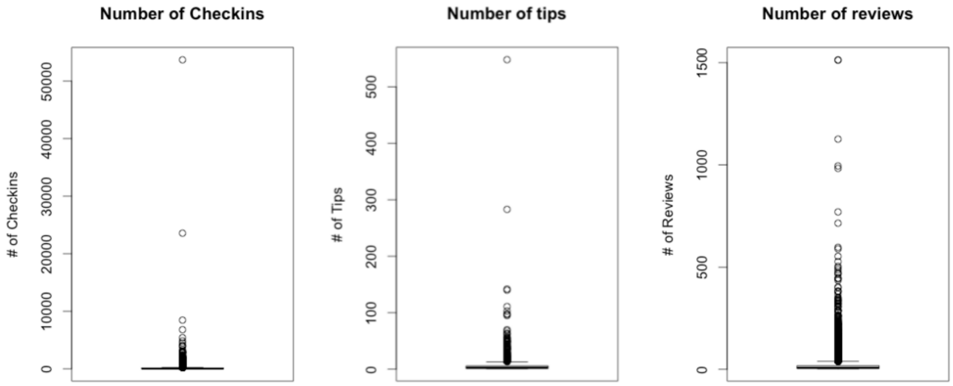
For further confirmation, we apply a linear correlation model to the same variables and plot them against Checkins as shown in Figure 3. Based on the strength of the linear association we determine that Tips are a stronger measure of popularity over time.



(a)

(b)

Figure 1: Time 0 corresponds to 2006-01-01. (a) shows a cumulative sum of users on Yelp over time. At time 36, after a short period of stalling there is a clear sustained growth. In (b) there is a parallel increase in number of tips and reviews over time specified in days.



[1,][2,][3,][4,][5,]	3, 12, 31, 92, 212	1, 1, 3, 6, 13	3, 4, 8, 18, 39
$N_{Outliers}$	1218	1181	1218
$N_{Observations}$	10000	1363	10000

Figure 2: Boxplots performed for the purposes of outlier detection. The first line of text under the graphs indicate the number of data points found in each respective segment of the boxplot. The sample size, and number of outliers detected for each variable are shown below in the two subsequent lines.

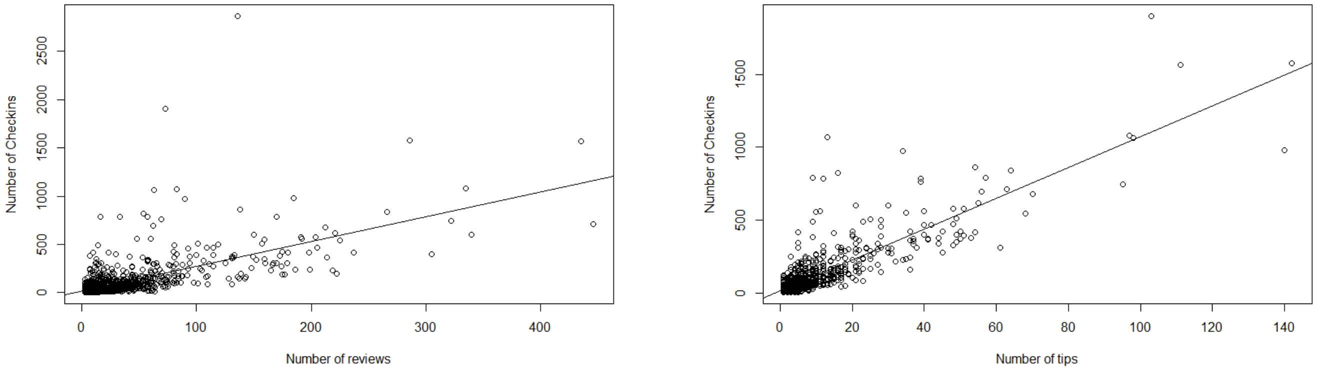
TABLE 1: Strength of linear association between variables

Variables	Pearson Correlation
Checkins and Reviews	0.65
Checkins and Tips	0.85
Checkins and Reviews + Tips	0.75

4.2. Selecting the right businesses

Having established Tips as our measure of popularity, for each business we plot them over time to look for significant positive trend changes over the history of the business such as in Figure 4.

The segmentation and split nature of the dataset can generate skewed results as for any given business we may be missing Reviews or Tips that are scattered over the rest of the 1.6M Reviews on record. Table 2 contains the characteristics of 11 businesses which fulfill our selection criteria, and which we used for statistical analysis.



(a)

(b)

Figure 3: Regression model applied to Reviews and Tips respectively in (a) and (b)

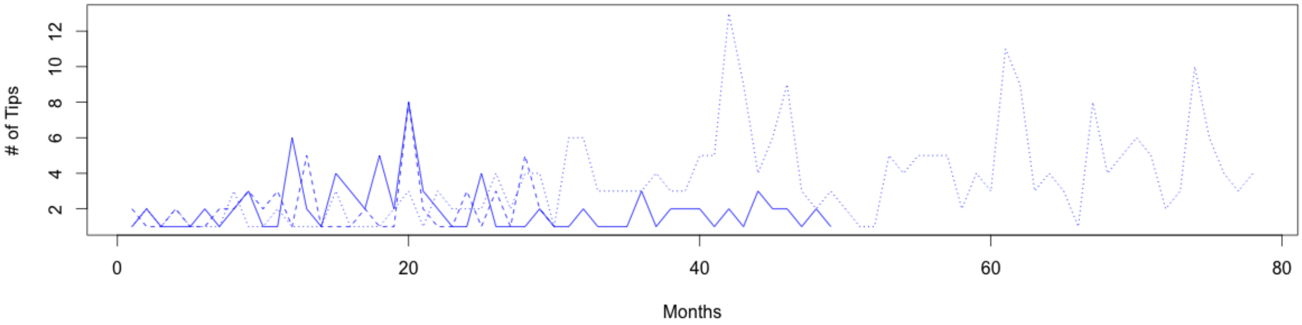


Figure 4: Tips profile of 3 selected businesses.

4.3. Social Network Analysis and Trendsetters Identification

For each business k, any user who has ever left a Review or a Tip is considered a node $v \in N_k$ in a directed graph $G_k(N_k, E_k)$; E_k is the set of all the edges (u, v) where $u, v \in N_k$. We search the network for TS by ranking every node in the network according to 4 methods: Indegree, Eigenvector centrality, Pagerank, and the Trendsetter algorithm published by Saez-Trumper et.al [4].

The first three are commonly used measures of node value within a network [12] and provide a reference frame for the performance of the TS algorithm on the Yelp social network. The TS rank $TS_k(v)$ of a node v in a network $G_k(N_k, E_k)$, is given by:

$$TS_k(v) = dD_k(v) + (1-d) \sum_{w \in In_{G_k(v)}} TS_k(w) I_k(w, v)$$

where: $0 \leq d \leq 1$ is the dampening factor, D_k is the probability distribution over all n_k, and $I_k(w, v)$ is the influence of w over node v. A more detailed elaboration of the TS ranking algorithm can be found in the work published by Saez-Trumper et al.

In this work we use a normal distribution with $D_k(v) = 1/|N_k|$ and set $d = 0.8$. We start with $TS_k(v) = 0.5$ for all $v \in N_k$ and iterate over them 15 times to let the $TS_k(v)$ values stabilize. Finally we consider both quantitative, and qualitative algorithm performance by evaluating the top 60 nodes for each ranking.

We measure quantitative performance by calculating what fraction of N_k interacted with a business k before peak popularity P_k: for each business we count the number of nodes v such that $v \in N_k$ and $P_k - T(v)_r < 0$ where $T(v)_r$ is the time at which node v from ranking r interacted with business k. We also measure algorithm qualitative performance by calculating the ratio of influenced friends $IF_k(v)$ by the top 3 nodes of each ranking, defined as the fraction of friends of v that interacted with business k after v.

TABLE 2: Characteristics of businesses that met selection parameters.

Business ID	Number of Reviews	Number of Tips	Nodes	Edges	Popularity peak
2X5G4Ujq0s4Wfn4TC7gX0g	208	68	131	636	2012-01-01
XLqnjlLYt0_q_NG7l_BpMA	61	103	53	230	2012-07-01
oCA2OZcd_Jo_ggVmUx3WVw	309	97	127	716	2012-03-01
mpDxBBGywUE6GRRKja3sBA	402	63	109	392	2012-11-01
Xo9Im4LmIhQrzJcO4R3ZbA	172	61	63	232	2013-06-01
45puCRQ6Vh_IIAy7kkfFDQ	155	57	78	468	2012-04-01
DlCtdbceo4YNSI53cCL2lg	203	55	95	430	2011-02-01
MwmXm48K2g2oTRe7XmssFw	169	140	144	1246	2011-12-01
Cp6JGY5YIRncTV_My9nf9g	192	70	51	214	2012-04-01
tb24fvNJfHhyKEXkKn12Xw	250	64	71	254	2013-02-01
McikHxxEqZ2X0joaRNKlaw	83	53	52	254	2012-01-01

5. Results

Figure 5 illustrates the number of users ranked as TS by each algorithm. Within the context of individual businesses, all algorithms return comparable number of users: the average standard deviation of algorithm performance across all businesses is in fact 3.18, with the minimum being 0 and the maximum being 6.4.

For a node to represent a Trendsetter however, it must be someone who interacted with the business before it hit peak popularity. Figure 6 shows how many of the users ranked by each algorithm satisfy this requirement. We can see that a comparable pattern develops along the plot, with TS and Eigenvector centrality following each other closely as well as Indegree with Pagerank.

Finally to evaluate qualitatively the inherent ability of users to influence their network, we calculate how many of their friends interact with a business after them. Figure 7 clearly shows that TS ranking was consistently and significantly better at identifying users who positively influenced their peers. This is in agreement with the Saez-Trumper et.al [4], confirming that users with high TS rank tend to influence more their social contacts than the users selected by other ranking algorithms. The poorest performance in this metric being that of Pagerank which was developed and used as ranking algorithm by Google.

6. Conclusion

Yelp is a commonly used platform to obtain information about businesses in the USA. Understanding the trends, preferences and connections between its users has an important value to businesses. In this paper, we show the application of different ranking algorithms to identify Trendsetters, and compare their relative performance.

We found that the studied algorithms had similar performances in finding who interacted with a business before its peak of popularity, yet the TS algorithm performs better when compared to Indegree, Eigenvector centrality and Pagerank in ranking users based on being early adopters, and able to propagate information widely.

This way, as for future work, we suggest to take into consideration the time variable (in this context the timestamps of Checkins) as a fundamental parameter in order to create directed graphs, and to model the TS equation for a more complete Social Network analysis.

References

[1] Yelp, *www.yelp.ca/about*, Web., 29 Nov. 2015

[2] Yelp, *yelp.ca/dataset_challenge*, Web., 29 Nov. 2015

[3] T. Zhang, and Y. Pan, *Yelp Challenge Project Report*, University of Washington Publication, 2014

[4] D.Saez-Trumper, *Finding Relevant People in Online Social Networks*, KDD'12, 2012

[5] E. Bakshy, B. Karrer, and L. Adamic., *Social influence and the diffusion of user-created content*, In Proc.of conf. on Electronic commerce, EC, 2009

[6] T. Jindal, *Finding local experts from Yelp dataset*, University of Illinois Urbana-Campaign, 2015

[7] J. Weng, E. Lim, J. Jiang, and Q. He., *Twitterrank: finding topic-sensitive influential twitterers*, In Proceedings of International Conference on Web search and data mining, WSDM 10,pages 261270, NY, USA, 2010. ACM. 10, 19, 30, 53

[8] T. Haveliwala.,*Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search*, IEEE Transactions on Knowledge and Data Engineering, 15:784796, 2003. 10

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi., *Measuring User Influence in Twitter: The Million Follower Fallacy*,In Proc.of the 4th Intl AAAI conf. on Weblogs and Social Media (ICWSM), 2010

[10] A. Anagnostopoulos, R. Kumar, and M. Mahdian., *Influence and correlation in social networks*, In Proc. of Intl conf. on Knowledge discovery and data mining, KDD, 2008

[11] G. Kossinets, J. Kleinberg, and D. Watts, *The structure of information pathways in a social communication network*, In Proc. Intl conf. on Knowledge discovery and data mining, KDD, 2008

[12] R. A. Hanneman, and M. Riddle, *Introduction to social network methods*, Riverside, CA: University of California, Riverside, 2005

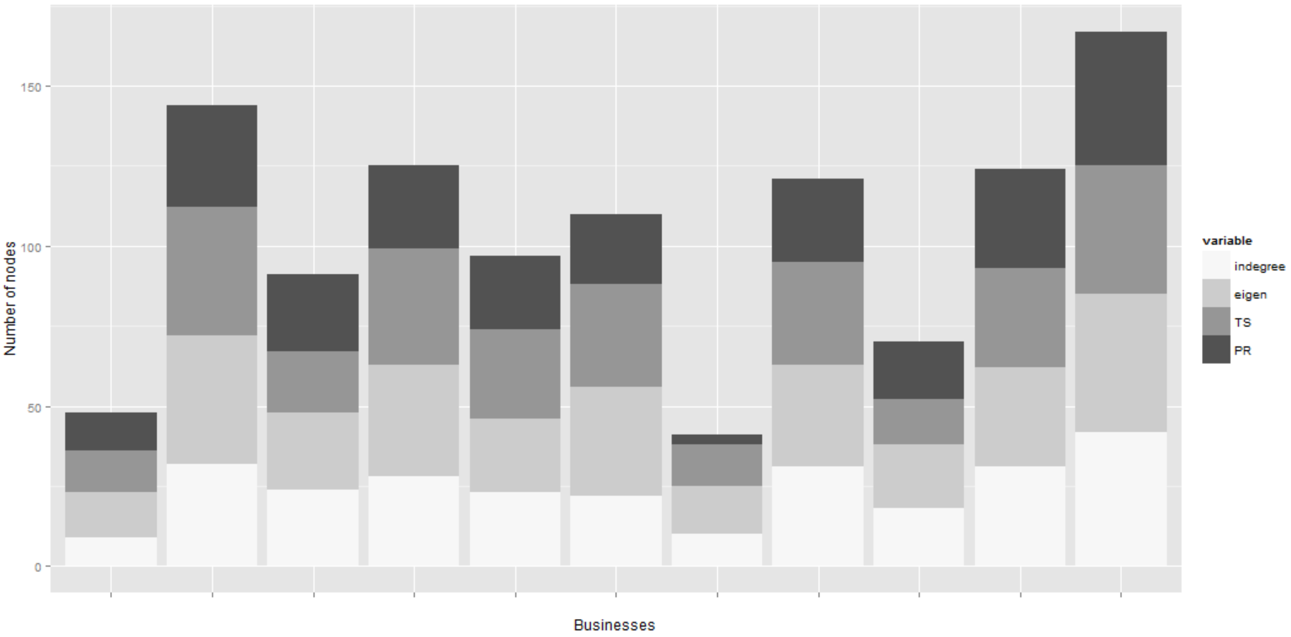


Figure 5: Quantitative performance per algorithm searching the network for Trendsetters.

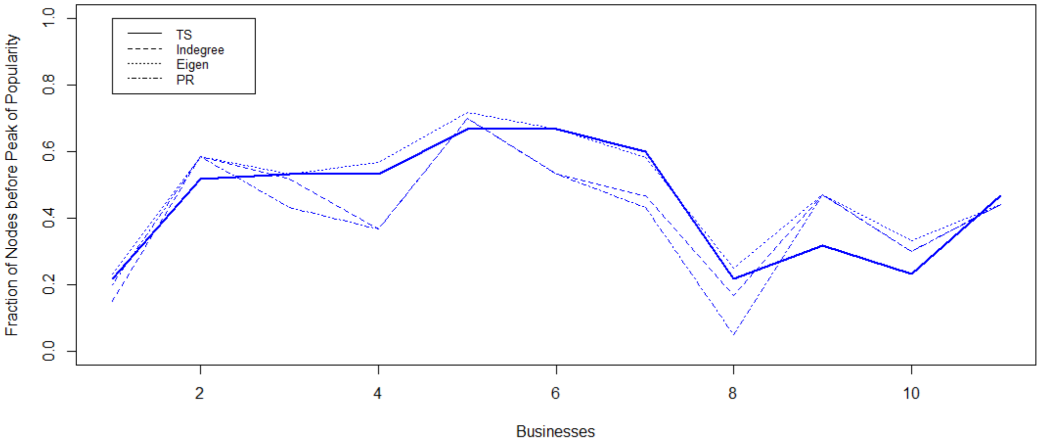


Figure 6: Percentage users of each ranking to interact with a business before it reaches its popularity peak.

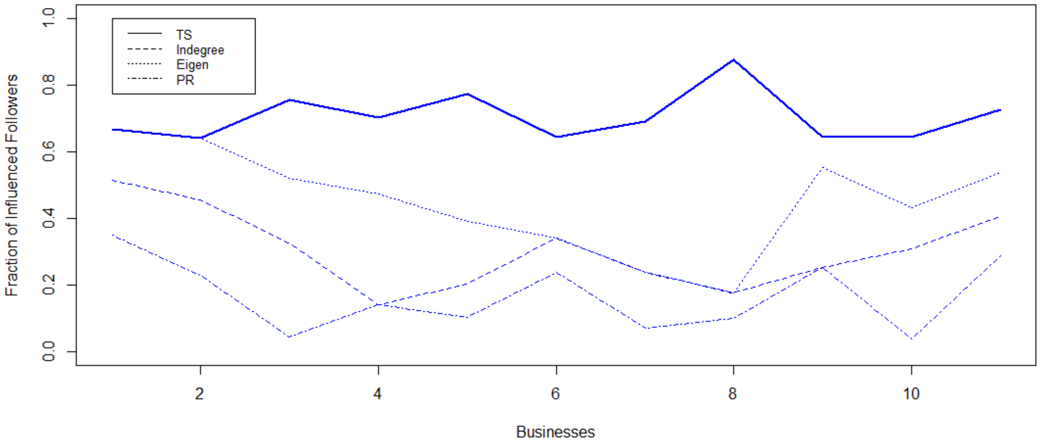


Figure 7: Average percentage of friends of the top-3 ranked nodes v that interacted with a business after v.