# Predicting Influential Papers
CS 224W Final Project
Autumn 2012

Shui Hu
Dilli Paudel

## Introduction

The problem of finding the most influential nodes in an influence diffusion network is actively studied in multiple fields. Much work has focused on the application of this problem to finding the optimal set of target customers in order to produce successful viral marketing campaigns. Others have attempted to find the most influential online websites or blogs by modeling the web as a diffusion network. Some work has also focused on finding the most influential set of academic papers on a particular topic. This third problem is relevant to researchers who often must read papers on new topics in order to complete their work. Since for most academic topics, more papers are published each year than any researcher can realistically read in his/her limited time, it is important for the researcher to read only the most influential papers in order to obtain an overview of the topic. Finding the most influential papers presents the additional challenge of deciding whether to read a paper soon after it is published. This challenge requires predicting the future influence of a paper. In our project, we will build a model to predict whether or not a paper will be influential based on its abstract, author, and citations. In order to find the necessary information, we will investigate the co-authorship and citation networks.

## Prior Work

A number of papers have studied the properties of co-authorship and citation networks, while others have studied influence in network diffusion. Most work on estimating the influence of nodes in a network have focused first on node-to-node influence, then used that to find the global influence.

One line of research attempts to learn the strength of microscopic influence of each node of a network on neighboring nodes by modeling network diffusion using the strength of node-to-node influence as a parameter, and then training the model on real data to obtain that influence. In “Unsupervised Prediction of Citation Influences,” Laura Dietz, Steffen Bickel, and Tobias Scheffer build a citation influence model that models each word in a paper. Their model learns a distribution of topics over each paper, the influence of each cited work, and the distribution of words per topic. By training the model from end-to-end, they obtain for any paper *d* and any paper *c*, the distribution P(*c* I *d*), which is the chance that *d* cites *c*–a measure of the influence of *c* on *d*. The influence of each paper on any other paper can be thought of as a first step towards quantifying the influence of any particular paper or author. Similarly, Amit Goyal et al use the General Threshold Model modified with time considerations to model the propagation of actions (like buying a phone) among members of an online social network (Goyal A, et al 2010). By training the model on the actions of the members of the network, they learn the probability that a person will be influenced by someone adjacent to him/her in the network into taking a particular action. In order to determine the macroscopic influence of each node of a network, one can combine the microscopic node-to-node influences of each node, e.g. by simply adding them.

One paper that does just that is "Mining Topic-level Influence in Heterogeneous Networks," which seeks to compute the macroscopic influence of each author on a given topic through the papers published (Liu L et al, 2010). This work uses two steps to compute the macro or global influence. First, they train a probabilistic model to obtain the influence of each node on each other node. Then, the influence is propagated through the entire network, leading to the global influence of each node. In the first step, the model simultaneously learns the node-to-node influence and the distribution of topics for each node. The distribution of topics specifies what topics the node is interested in. The probability that a certain node influences another node is parametrized in the model, serving as the node-to-node influence. This model uses the similarity of topic distribution between nodes as evidence of influence, which allows it to learn the node-to-node influence parameters.

Network diffusion is often affected by a combination of the network itself and external influences. For example, citations to a paper may suddenly increase when the industry finds a use for the topic covered by that paper, but the events happening in industry are not captured by the networks of co-authorships and citations. "Influence Diffusion and External Influence in Networks" builds a model that incorporates both network and external influences in describing diffusion (Myers S. et al, 2012). This model looks at the probability that each node is exposed to internal and external influences and uses those learned probabilities to determine the overall influence of external sources versus internal network sources. It found that for tweets on Twitter, 29% of URLs posted were due to external sources, implying that external influence is important to consider when computing the overall influence of each node.

Another approach to finding the most influential nodes in a network is more direct. David Kempe et al prove that standard hill-climbing approximation algorithms for computing the most influential sets of nodes in graphs following the Independent Cascade or Linear Threshold Models provide provably accurate results (Kempe D. et al, 2003). The Independent Cascade Model assumes that each node has a some particular probability of propagating an action to each of its neighbors, while the Linear Threshold Model assumes that each node activates (i.e. propagation reaches it) if and only if the number of its neighbors that has activated exceeds some threshold. These two relatively simple models are commonly used to model diffusion.

Still another approach to estimating the influence of authors can be found in "Identifying the Influential Bloggers in a Community" (Agarwal N et al, 2008). This paper uses the influence of blog written by a blogger to estimate the influence of the blogger him/herself, the intuition being that the influence of a blogger's blogs determines his/her influence. The influence of a blog is found to be a weighted sum of the length of the blog, the number of comments, and the number of outgoing links subtracted from the number of incoming links. The difference between the links is used because they found that influential blogs are cited often by other blogs *and* contain more original content, which reduces their number of citations. These are insights that may also apply to academic papers.

Finally, two papers also discuss the properties of academic co-authorship networks. Leskovec et al, finds that contrary to most social networks, co-authorship networks grow denser and their diameters shrink as time progresses (Leskovec et al, 2005). They believe that this is due to what they call the Forest Fire Model, which essentially stipulates that new authors tend to collaborate with one established author, then branch out to his/her previous collaborators and so on, reducing the distance between nodes over time. Earlier work by Barabasi et al shows that the degree distribution follows the Power Law which is consistent with the Forest Fire Model (Barabasi et al, 2002).

**Problem Definition**

Given a set of papers, their network of citations, the paper abstracts, and the author lists for each paper, we will predict whether a particular recently published paper will be influential in the future. The paper we investigate must have been published at a time that is not too distant from the publication times of the papers in the citation network. Influence of a paper is an ambiguous notion, but we will use the standard academic measure of the number of citations that a paper receives. The more citations a paper receives, the more influential it is. This measure is both easy to compute given a citation network, and is well-regarded by paper-writing academics. We use a threshold $T$ such that a paper is classified as influential (class 1) if and only if it is cited by at least $T$ other papers.

**Data**

We use data from the arXiv high-energy theoretical physics database containing paper submitted between 1992 and 2003, inclusive. This data divided into training and test data sets. Papers submitted between 1992 and 1997, inclusive, make up the training data, while those submitted between 1998 and 2000, inclusive, make up the test data. Papers submitted between 2001 and 2003 are not included in either set, because we want to give each paper at least a three year window to collect citations.

*Labeling the Data*

Assigning labels of influential and non-influential to the papers is not straightforward because the data does not have such labels to begin with. Unfortunately, it is not obvious what value of $T$ leads to the most logical labeling. In order to determine the threshold $T$, we find the $f_t$, the number of papers that are cited exactly $t$ times. Then,

$$T = argmax_T\ [Var(\mu_0, \mu_1) - Var(\{f_t \mid t \geq T\}) - Var(\{f_t \mid t < T\})]$$ where

$$\mu_0 = avg(f_t \mid t < T\})$$

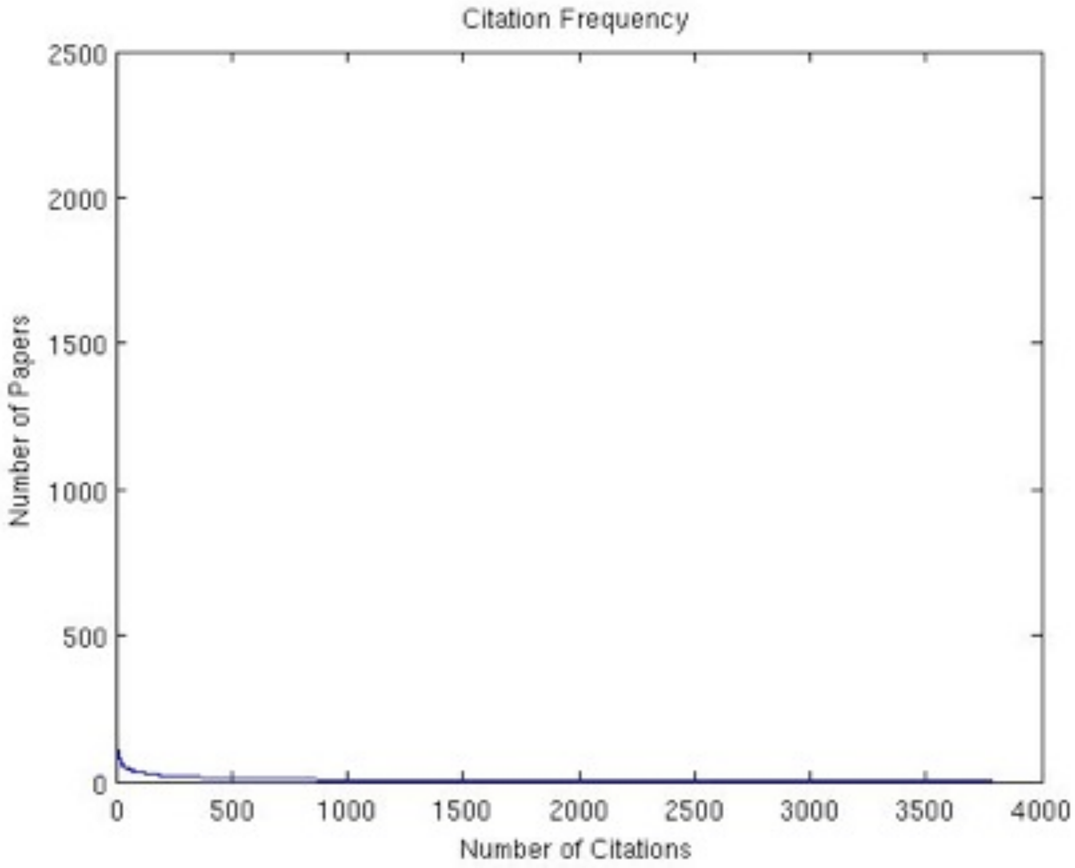
$$\mu_1 = avg(f_t \mid t \geq T\})$$

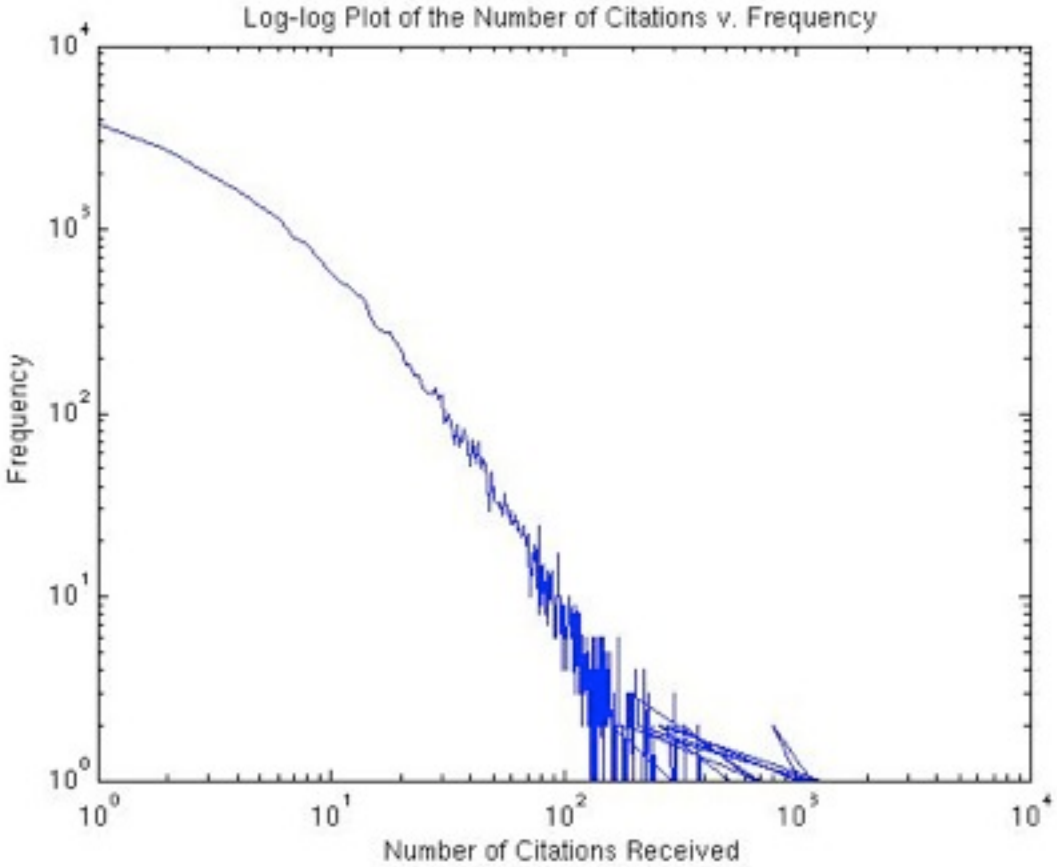
Note that this value of *T* maximizes the inter-class variance while minimizing the intra-class variance. As a result, *T* divides the data into two sets, one of which contains papers that clearly receive more citations than those in the other. By labeling papers using such a value of *T*, we capture an intuitive sense of influence among the papers. For our data, this leads to *T* = 10 and 29.7% of the papers being labeled influential. Formally, given a paper *p* cited $t_p$ times, the ground-truth label of *p* is

$y^{(p)} = 1\{ t_p \geq T \}$ (note: class 1 corresponds to influential, class 0 corresponds to non-influential)

**Network Properties**

The citation network, consisting of nodes that correspond to papers and edges that run from node *i* to node *j* if and only if paper *i* cites paper *j*, does not conform to many of the random network models studied in class. With 27,770 nodes, the diameter is 126, so the network is not likely to be a small-world random graph. Furthermore, the distribution of node in-degree (i.e. the number of citations that a paper receives) is more like an exponential curve than a power curve, which suggests that the preferred attachment model does not apply either.





As one can see in the above log-log plot of frequency of received citations, the curve on a logarithmic scale is not linear and has little weight in the tail. Thus, it is possible that the distribution of the number of citations received by a paper follows an exponential law since exponential curves also have light tails and are concave down on a log-log plot. The difficulty of fitting the citation network to a general model is not surprising, since the main motivation for an author to cite another paper is the relevance of that paper to his/her own paper, but this motivation is not represented well in the network. Without a clear model to use for this network, we default to building our own model using features of the network and abstracts. The works of Agarwal et al. and Yogamata et al. show that using features in simple linear models tends to work in such problems.

Additionally, we also found that papers tend to cite other papers that were written recently. Few citations are made more than three years after a paper is published. This is the reason for using the year 2000 as the final year of papers to be considered for the test set. This phenomenon also suggests that we should weight features based on recency of the paper the feature comes from when classifying.

## Algorithm

We use logistic regression and the following features to classify each paper.

Logistic Regression:

Let $y^{(i)}$ and $x^{(i)}$ be the ground-truth label and feature vector of paper *i*.

$p(y^{(i)} = 1 \mid x^{(i)} ; w, b) = 1 / (1 + exp(-(w^T x + b)))$

The logistic regression model is trained using maximum log likelihood estimation over the training data. We use stochastic gradient descent to speed up the learning process.

Features:

*Basic Features*

***1)*** Number of citations in current paper

***2)*** Weighted sum of the number of people that each author of the current paper has coauthored with so far. Since the first person on the list of authors is the most important for the paper, the weight decreases as we progress through the list.

***3)*** Number of authors for the current paper.

***4)*** Weighted sum of the number of citations received by the authors so far in their previous papers. The weights here are same as for feature 2.

***5)*** Length of abstract.

*PageRank Features*

***6)*** The weighted sum of the PageRank scores of the papers that the current paper cites. The PageRank is computed for all papers published so far and assumes that each paper contributes PageRank to papers that it cites. The weight used for a paper is the reciprocal of the number of the number of papers that cite it.

***7)*** The weighted sum of the PageRank scores of the authors of the current paper. This PageRank is computed for the co-authorship network where each node corresponds to an author and if author *a* cites a paper written by author *b*, a directed edge runs from *a* to *b*. The weighting is the same as for feature 2.

*Text Features*

***8)*** Vector containing the TFIDF of the terms in the title. The dimension of the vector is the size of the vocabulary. To limit the dimension of the vector, the vocabulary to limited to words that are relatively common, but not the most common (to avoid stopwords). Pruning of the vocabulary leaves a vector of at most 1000 elements which corresponds to words that occur often enough to lead to correlations between papers that contain them.

***9)*** Vector containing the TFIDF of the terms in the abstract. The vector is the same format as for feature 8 and the the vocabulary for abstracts has been similarly pruned.

These features are concatenated for each paper to yield a vector representation $x^{(i)}$ of dimension around 1000. The basic features are motivated by the intuition that people cite other papers based on the reputation of the authors, the professional networks of the authors, the record of citations of the authors, and the accessibility of the paper (shorter abstracts catch reader's attention more often than long abstracts, leading to more citations). On the other hand, the PageRank features are motivated by the fact that high PageRank scores correlate strongly with large numbers of citations. In fact, PageRank is arguably just as good of a measure of influence as the number of citations. Unfortunately, when a paper has just been published an no one has cited it yet, one cannot obtain an accurate eventual PageRank of the paper, so our features 6 and 7 approximate the true PageRank by going backwards through the PageRank computation. We assume that after a while when the true PageRank of the current paper has been calculated, the PageRank scores of the papers it cited will not have changed very much. Since the PageRank score of those papers come from the PageRank scores of the papers that cited them, we assume that those incoming citations each contributed evenly to the PageRank, leading us to the computation we have for the current paper in features 6 and 7. The reason that we compute PageRanks for both papers and authors is that in practice, citations are correlated with both the content of the paper being cited and its author (due to reputation or other people factors). Finally, the text features attempt to capture the important content of each paper by looking at the important words in the abstract and title.

## Results

Test Set Results

| Features | Precision | Recall | F1 |
|---|---|---|---|
| Basic Features | 0.4417 | 0.9968 | 0.61214 |
| Basic Features + PageRanks | 0.48003 | 0.99472 | 0.64756 |
| Basic Features + PageRanks + TFIDFs | 0.48109 | 0.99386 | 0.6618 |

Due to the uneven number of class 1 and class 0 instances, it makes more sense to examine the precision, recall, and F1 scores for class 1 (influential). Though displayed here, the train scores were very close to the test scores even without applying regularization when training the logistic regression model, so the classifier generalizes very well. As the numbers show, it is relatively easy to obtain very strong recall, but precision lags behind due to the lack of features that distinguish well between the two classes. The above results are found to be insensitive to different parameter initializations as long as the weight vector *w* is initialized with all components close to zero.

## Analysis

*Low Precision*

The results indicate that precision hovers around 50% which means around 40% of class 0 papers are misclassified. This is due to lack of discriminative power by the basic features combined with the inaccuracy of the PageRank approximation. When clustering the $x^{(i)}$ by Euclidean distance, we found that the misclassified non-influential papers are also grouped incorrectly with the influential papers. Our PageRank approximation suffered whenever a non-influential paper cites an influential paper–it receives a portion of the high PageRank even though it does not contribute such a large amount. Citing an influential paper does not make one's own paper influential it appears. Thus, the PageRank approximation also contributed to the problem. A better approximation would have to mitigate this problem.

*Inconsequence of Text Features*

Text features, which other researchers have found to be very helpful in predictions, are not very helpful here. This is due to the small vocabulary after pruning which left many TFIDF vectors full of zeros. Better stemming of the words beforehand would have reduced the size of the vocabulary without reducing the effectiveness of the feature, but the many irregularities in the text made consistent stemming difficult.

*Abstract Length*

Contrary to expectation, the weight for abstract length was a large positive number in our training runs, suggesting that long abstracts correlate with more influence for the paper in the long run. Perhaps, longer abstracts usually correspond to papers with more important content.

Given the above findings, it appears that one could produce comparable performance by just using the basic features which are much faster to produce with larger datasets compared to computing the PageRank scores or TFIDFs. Furthermore, at least with the above data, it appears that network data alone (basic features capture essentially just network information) without content of the papers can be used to recognize influential papers, but it cannot discriminate well between influential and non-influential papers.

## Conclusion

Using a logistic regression models with appropriate features selected from the citation and co-authorship networks as well as some PageRank and textual information, one can predict the future influence of a paper with high recall, but low precision. The low precision is due to the lack of distinguishing features that separate influential from non-influential papers. In practice, due to the time it takes to train PageRank on large datasets (not a problem for this dataset, but the arXiv set is small), one can obtain similar results by using just the features from the networks.

## Works Cited

A. Goyal, F. Bonchi, L.V.S. Lakshmanan. *Learning Influence Probabilities in Social Networks*. In Proc. WSDM, 2010.

A.L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek. *Evolution of the Social Network of Scientific* Collaborators*.* Physica A: Statistical, 2002.

D. Kempe, J. Kleinberg, E. Tardos. *Maximizing the Spread of Influence in a Social Network*. In Proc. KDD 2003.

D. Yogamata, Heilman, M, O'Connor, B, and C. Dyer. *Predicting a Scientific Community's Response to an Article.* EMNLP, 2011.

J. Leskovec, J. Kleinberg, C. Faloutsos. *Graph Evolution: Densification and Shrinking Diameters*. ACM TKDD, 2007.

L. Dietz, S. Bickel, and T. Scheffer. *Unsupervised Prediction of Citation Influences*. ICML , 2007.

L. Liu, J. Tang, J. Han, M. Jiang, S. Yang. *Mining Topic-level Influence in Heterogeneous Networks*. CIKM, 2010.

N. Agarwal, H. Liu, L. Tang, P. Yu. *Identifying the Influential Bloggers in a Community*. WSDM, 2008.

S. Myers, C. Zhu, J. Leskovec. *Information Diffusion and External Influence in Networks*. In Proc. KDD, 2012.