1 Introduction

In this presentation, we discuss the paper “Can Cascades be Predicted?” by Cheng, Adamic, Dow, Kleinberg, and Leskovec, in the broader context of the recent work done on information sharing in social networks.

1.1 Why Do We Study Cascades?

Information disseminates through social media to be shared by a range of people. Sometimes, such content can be shared by millions of people belonging to a wide range of age groups, backgrounds, political views, nationalities, religious views, genders, etc.

A famous instance is a picture of a receipt taken by a customer at a hunting and fishing chain store called Cabela. The receipt listed a 2.3% “medical excise tax,” which was supposedly levied as part of the Affordable Care Act.

This was seen as evidence that there is a hidden tax that consumers are paying for the ACA. The store later cleared this up as being a “software glitch” on their part. This tax is supposed to be payed by merchants and importers, and not consumers, and it is only applied to medical equipments not fishing equipments. The store supposedly
charged the consumers by accident. Nonetheless, the picture of the receipt was reshared over thousands of times even in the first month that it was posted.

Since such a cascade of resharing is one of the primary ways in which people discover and consume information, it is important to understand what determines the popularity level of such posts. This can be useful in various domains including sales and advertising, formulating and testing sociological theories, and advocating for different causes.

Before presenting the results of this paper, it is important to formalize the notion of cascades on a network. As usual, a social network can be modeled by a finite graph $G$ where the nodes represent people (or company pages) and the edges signify “friendships” or “following”. Content is produced at one of these nodes and can get reshared by its neighbors, which can then pass it on to their neighbors and so forth. This starts a cascade, whose graph we denote by $\hat{G}$. This is a rooted tree that encodes how information passes through the network. We denote $\hat{G}$’s induced subgraph by $G'$. Thus, $\hat{G} \subseteq G' \subseteq G$.

**1.2 Overarching Goals**

To understand how content is reshared, we focus on two aspects of the graph $\hat{G}$—its size and structure. That is, we care about the growth and virality of the graph.

The former is easy to quantify and it is a direct measure of the quantity of influence in the network. Virality can be measured in different ways (one of which we will discuss
in more detail in the next section). While this measure is more subtle, it is equally important since it measures the impact and nature of the influence. Cases when $\hat{G}$ is the star graph with content initiated at the center is called a broadcast model. This is one of the least viral graphs. Cases where the content is able to reach and get passed around different “components” of the network are, on the other hand, more viral. They are seen as more organic and useful in cases when we care not only about how many people a content reaches but whether it reaches different groups of people. Take, for instance, the public service announcement called “It’s on US” released by the Obama administration to end sexual assault.

Attempts to understand the growth and virality of such graphs have faced two main obstacles:

1. Large cascades are rare.
2. The eventual scope of a cascade might be an inherently unpredictable property.

The first is intuitive; indeed, very few pictures and public posts go viral on Facebook. This concern is also easy to quantify. We can take the set of all public posts on Facebook and see the number of times they were reshared to note that viral posts are outliers.

The second has much stronger implications but is also harder to quantify. So, instead of trying to prove or disprove this, we shift our focus to how much we can say about the growth and trajectory of cascades. In the process, we will also point out which features of these cascades are helpful in this prediction task.

Such a prediction problem is complex. Assume you are given the initial few stages of a cascade and asked to predict how large it will grow and what its structure is likely to look like. This is a pathological task since most cascades are small and do not have deep structures. If we instead consider just large cascades, then this creates an artificial setting that is unrepresentative of how content spreads in social networks. As such, a lot of cascade prediction attempts have been been prone to the above issues and a robust formulation of this problem is an open problem in the field of social networks.
1.3 Previous Attempts
We list the approaches of most related papers and issues associated with them:

• Focus on fixed values such as predicting whether a given individual will reshare a specific content, predicting whether a cascade will increase in size by a particular amount, and so on.

• Focus on aggregate activity such as predicting the number of upvotes, total daily hashtag use, etc. These above two approaches are problematic in that they focus on groups of already non-trivial sizes or the questions are empty in that most content is not large and viral so will not be reshared by most individuals.

• Focus on prediction after observing a cascade for a given fixed time-frame. This can be problematic for slowly but persistently growing cascades.

• Considering the cascade growth as a regression problem or a binary classification problem with large bucket sizes. These can however be biased towards very large but extremely rare cascades.

1.4 Contributions
The main contribution of the paper is as follows: Predicting properties of cascades is difficult because most of them are small. However, asking nuanced questions can lead to powerful predictions.

More precisely, the paper proposes a novel way of formulating the prediction problem, which does not suffer from the biases induced by the scarcity of large cascades. This new prediction problem can be applied to both size and structure prediction and leads to the following observations:

• Size and structure can be predicted with a reasonably high accuracy.

• The predictability of a cascade increases as you observe more of it.

• User initiated cascades are harder to predict than cascades initiated by organizations (companies, celebrities, etc.).

• The relative importance of the different features for the prediction problem evolves over time.

2 Methodology
2.1 Problem formulation
Prediction problem Let $f(k)$ be the median size of all the cascades with at least $k$ reshares. We can define the following classification problem: given the first $k$ reshares, predict whether or not the cascade will eventually have a size larger than $f(k)$. It is easy to see that this classification problem is “balanced,” in the following sense: by definition,
exactly 50% of the cascades with at least \( k \)reshares will eventually reach a size at least \( f(k) \).

Rather than being a fixed value \( k \) is a free parameter defining a series of classification problems, one for each \( k \). This allows us to study how the predictability of a cascade evolves as more of it is observed. This is particularly relevant in the context of information cascades in social networks, where you could imagine an entity monitoring the progress of a cascade over time and trying to make predictions as the cascade grows.

**Wiener index** As mentioned before, the virality of a cascade is also an important factor. However, in order to define a classification problem for virality akin to the one with size, we need a way to quantify the virality of a cascade. Specifically, we need an index (a real number) capturing our intuitive idea of a viral cascade.

![Figure 5: Two cascade shapes. A good virality index should be low in the broadcast diffusion model (a) and high in the tree diffusion model (b). For the Wiener index, we have that \( \nu(T) \approx 2 \) for (a) and \( \nu(T) = O(\log n) \) for (b).](image)

A good virality index should be low (constant) in the broadcast case and high for the complete binary tree. Simple such indices which satisfy those properties are the maximum depth or the average depth of vertices to the root. However, it is easy to find simple examples where these indices fail. It has been recently suggested that the Wiener index is a better index of virality. The Wiener index is simply defined as the average distance between pairs of vertices. Formally, the Wiener index \( \nu(G) \) of a graph \( G \) is defined by:

\[
\nu(G) = \frac{1}{n(n-1)} \sum_{i \in G} \sum_{j \in G} d_{ij}
\]

where \( d_{ij} \) denotes the shortest path distance between \( i \) and \( j \).

We note however that the Wiener index is not perfect either. For example, a path with \( n \) vertices has a virality index \( O(n) \) even though it does not correspond to the intuitive idea of virality. Experimental evidence seems to indicate that the examples for which the Wiener index fails are not as commonly encountered in social networks as examples for which other virality indices fail.
2.2 Dataset

The dataset used consists of public pictures on Facebook from June 2013 each of which have been reshared at least 5 times. Specifically, the reshares of 150572 photos were tracked for a period of 28 days. (The photos were reshared a total of over 9 million times over this period.) For each of these photos, using reshares, clicks and views data, it is possible to reconstruct the information cascade. In particular, some care must be taken when attaching a node to its parent. User $C$ could reshare from user $A$ even though she was originally exposed to the content through user $B$. Using clicks and views can help disambiguate those situations where reshares would only lead to believe the cascade to be more shallow than it really is.

Having access to this dataset, the first thing to look at is how the cascade size and virality are distributed. This can be seen in Figure 6.

An important distinction should be made between user initiated cascades and page initiated cascades. On Facebook, pages are everything which are not “real-life” people: institutions, celebrities, etc. The reason for this distinction is that user-initiated cascades are generally considered to be more organic, hence more likely to become viral, whereas page-initiated cascades, even though they might be seen by more people (pages will typically have many followers) tend to be less viral.

Second, we see that the cascade size distribution is approximately a line in log-log plot. This indicates a power-law distribution. By fitting a power-law distribution to the data, we obtain a power-law exponent $\alpha \simeq 2.1$. We note that this experimental evidence supports some already existing theoretical models for cascades for which a formal derivation of the cascade distribution is tractable and is indeed a power-law distribution.

2.3 Classification

It is easy to see that for a power-law distribution of exponent approximately 2, the median of the distribution conditioned on being larger than $k$ is exactly $2k$. As a consequence, the classification problem now becomes:
Given that \( k \) reshares have been observed, is the cascade going to double in size, i.e. reach at least \( 2k \) reshares.

The methodology used by the authors is to treat this problem as a standard classification problem: extract features from the first \( k \) reshares and apply off-the-shelf classification algorithms (linear/logistic regression, SVMs, Naive Bayes, random trees, etc.) The authors note that most of the algorithms perform very similarly on their dataset. Hence, most of the focus is put on designing good features and understanding their relative importance.

The features they consider in this paper are: content, root (original poster), resharer, structural, and temporal features. We refer to the original paper for a list and detailed description of each of these.

3 Results

3.1 Predicting Cascade Growth

The first result is a logistic regression performance using the features mentioned in the previous section.

![Figure 7: Accuracy Level of the Different Features](image)

We compare these results to the baseline, which is to randomly guess whether a cascade is going to double in size or not with equal probability. Their methodology achieves notably stronger results with classification accuracy of 0.795. Furthermore, if this was relaxed to guess whether a cascade was in the top or bottom quartile (that is, top 25 or bottom 25 percent), they are able to predict with an accuracy level of 0.926.

It is important to notice how the different distinct features do individually. Temporal features come out on top with accuracy of 0.780, which is only 0.015 percent away from considering all the features. This hints at relative significance of temporal features. However, they are able to do well without the temporal features, gaining an accuracy of 0.722. The other features also give strong results on their own compared to the baseline.

Since \( k \) is a non-fixed parameter, a natural question to ask is whether the results get stronger or weaker as \( k \) grows. On the one hand, small values of \( k \) mean that they have
observed less of the data, making prediction harder. On the other hand, higher values of \( k \) might introduce more noise into the data and also require predicting further into the future. As it turns out, the accuracy of their results increases with growing value of \( k \), although it is worthwhile to notice that the mean accuracy only increases by 0.03\% as \( k \) varies from 5 to 100. An interesting observation is that the content seems not to matter as much, especially with higher values of \( k \).

![Figure 8: Changes in Accuracy with Growing \( k \)](image)

This methodology also allows us to ask how the significance of each of the features is affected as the value of \( k \) increases. We focus on two features at the opposite ends of the spectrum, although the paper presents results for each of the different features.

We have already noticed that content might not be a strong predicting factor. When they compute the value of the feature after observing the first \( k \) reshares and plot that against the correlation coefficient of the feature value with the log-transformed number of reshares, they see that the actual content tends to zero for higher values of \( k \). The exception of content features are whether the picture has a caption and whether the caption is in English. This begs the question whether anything can be viral.

![Figure 9: Changes in Content Importance](image)

Temporal features, on the other hand remain consistently important, or even increase in importance, with increasing values of \( k \). Take, for instance, the number of users who
saw the first $k - 1$ reshares, until the last reshare was posted. This steadily increases in importance and more than doubles the accuracy from 5 reshares to 50.

Figure 10: Changes in Temporal Importance

3.2 Predicting Cascade Structure

The results on cascade structure predictability are very similar to the ones on size predictability. We refer to the original paper for a full description and only highlight the following points:

- Cascade virality can be well predicted, the reported accuracy is 0.725.
- Structural and temporal features are again the two most important features.
- The relative importance of these two group of features compared to others becomes more important as $k$ increases.
- Cascade virality predictability increases with $k$.
- Page-initiated cascades are easier to predict.

4 Discussion and Open Problems

4.1 Strengths of Paper

Here are some strengths of the paper that we want to highlight:

1. The dataset is new and interesting. The data analysis performed on this dataset is thorough and reveals many important aspects of information cascades in real social networks.

2. This new formulation of the growth prediction problem is very elegant since it is naturally balanced.

3. Having $k$ as a free parameter allows us to work around a big issue, which is that large cascades are very rare but they can bias analysis due to their size. Moreover, focusing on fixed analysis is not the right question to ask since most cascades are small and thus guessing the ultimate size of a cascade or predicting whether a specific individual will participate in the cascade is an almost empty question.
4. The fact that the classification problem can be redefined as we observe more re-shares creates a new spectrum of questions related to how predictability and different properties of cascades change over time.

4.2 Open Problems

Here is a list of unresolved issues and open problems:

1. We have seen above that while temporal features are the most important single feature, they are able to do well without them. What about the other features, and especially content features? Are we able to completely ignore any of the other features and get an accuracy close to what we would get when considering the remaining features?

2. The study is performed on publicly shared Facebook pictures over a short period of time during June 2013. It is difficult to say how representative publicly shared Facebook pictures are of how information disseminates in social networks. For instance, Twitter might inherently be more conducive to resharing and disseminating information. How would this methodology hold in this domain or over different time-frames? In particular, we have noticed that content of the photo becomes less important with higher reshares with the exception of whether the picture has a caption and whether the caption is in English. This hints at the possibility that text based cascades might not have the same result as picture based ones!

3. A related issue is that this study focuses on photos that are reshared, but does not take into account ‘liking’ or ‘commenting’ on these photos. While these are arguably smaller measures of impact compared to resharing, it is not clear that they should be completely dismissed, especially as on Facebook ‘liking’ public content can result in one’s neighbors to be exposed to it that they otherwise might not have.

4. Conclusions about the relative uselessness of content features compared to other features are weak. It is hard to extract meaningful features from the content, and the fact that the chosen features perform badly is more an indicator that these features are badly designed than the content having no impact on cascade growth or virality.

5. Is Wiener index the best measure of virality?

6. If original structure of the first five nodes is so important, as shown in the paper, why was it not used as a classification feature?

7. It would also be interesting to see consequences of the size and especially virality of cascades. For instance, are they correlated with higher sales or increase in activism?