Repin This: Analyzing Influence Patterns in Pinterest

Michael Chen Sam Kim Lucas Lemanowicz

CS 224W, Stanford University

Abstract

In this paper, we construct graph inference models and relevant statistics to investigate how Pinterest disseminates information. Pinterest is unique in how information spreads in the form of "repinning". We identify some key characteristics of this network. We compare two methods of constructing the underlying network behind Pinterest. After modeling the relationship between pins and boards, we observe how pins propagate across boards. We then identify what makes boards influential on Pinterest. Finally, we perform a temporal analysis on the output of our influence models to learn how influence changes over time. We believe this could provide researchers and marketers insight into the characteristics of influential boards as well as strategies for advertising.

1 Introduction

There are many ways to be influential in a social network: for example, one could initiate a new trend, amplify new trends, or act as the deciding factor in which trends continue. We identify several ways to quantify influence in Section 2. Pinterest presents an unusual social network structure centered around boards; in order to study their influence, we identify boards that initiate the propagation of information and trace the edges involved in the respective cascades.

On Pinterest, users generally repin from the homepage or pin directly from other websites. Typically, users do not care which board they borrow a pin from [1], so we expect our unbiased graph models to effectively represent the true network without having to consider the pin's content. However, because the underlying network is largely unobserved (we know when pins were repinned but we do not know the originating board), one challenge is to reconstruct the network before we analyze how influence propagates.

We compare results from using a board latency model as well as a network inference model to estimate the underlying network. Once we have a graph model of repins, we analyze this graph for influencers, cascades, and bridges. Our project objective is to identify and characterize 'influential' boards in various definitions of 'influential'.

Being influential on Pinterest is important for marketers because the social network lends itself to being an excellent driver of e-commerce sales.

2 Related Work

We survey several works analyzing how information disseminates in social networks. We build upon the models and algorithms these authors develop as ways to analyze the Pinterest network. Furthermore, we looked into several papers looking specifically at Pinterest to better understand what typical user behavior entails.

2.1 Quantifying Influence

First, we need a way to quantify influence on Pinterest in order to identify which boards are more influential. Bakshy *et al.* [2] quantify the influence of individual "influencers" as well as tweets in their Twitter dataset by identifying users who "seed" content, meaning they are the first one to post tweets. They measure different levels of influence depending on how many users a given user follows posted the same URL. For example, if C follows A and B, and C posts the same URL after A and B have both posted it, the level of influence is split among A and B. These directed edges of influence form an influence tree, referred to as "cascades".

Their model resulted in three key indicators of influence: past performance, number of followers, and activity level. However, their model does not take time into consideration, even though posts close in time are more likely to be correlated than posts far apart in time. We plan to factor in the time between repins in our work.

2.2 Inferring Networks of Influence

Before we are able to quantify influence and analyze the graph for influencers and cascades, the network has to initially be inferred. Gomez *et al.* [3] proposes a scalable algorithm that infers the network based on times online blogs and articles share hyperlinks between different news sources. They formulate a generative probabilistic model of how, on a fixed hypothetical network, cascades spread as directed trees through the network. Although under that model, the network inference problem is intractable, they develop a tractable approximation that show nearly optimal results. Because our problem formulation for creating a network falls in the category similar to that of Gomez, we adapt their proposed algorithm to infer a network on the Pinterest dataset and proceed to analyze the resulting graph for influence.

2.3 Information Diffusion and Cascades

Next, we want to observe how information spreads on Pinterest. Gruhl *et al.* [4] derive a mathematical model of information propagating through Blogspace by analyzing individual topics and seeing how bloggers influence one another. They establish a confidence for each user by measuring how many standard deviations above or below the mean time a user posts about a topic. They then establish a "copy probability", which is how likely they think that a user B copied off of A after reading user A's blog. The probability increases as the number of cases of A posting about a topic before B increases. Then, to analyze communication across Blogspace, they develop the transmission graph model, where edges are more likely to appear if there is a high copy probability to substantiate it. The transmission graph model will help us turn the nested bipartite graph in the Pinterest data into a graph measuring how information flows between boards.

2.4 Learning Influence Probabilities

We also look into other ways to build a graph-based model to measure influence. Goyal *et al.* [5] use a Flickr dataset (1.3 million nodes, 40 million edges, 35 million actions of which 300 thousand are distinct) to explore ways of calculating influence scores from a graph of user actions and interactions. Specifically, they compare three model types that are applicable to influence analysis: static models, continuous time models, and discrete time models.

Static models ignore the time component of actions and assume that a user is equally likely to be influenced by another user's action regardless of how long ago the action was taken. However, this model is not particularly helpful to us, since we want to consider time between actions.

Continuous Time models are predicated on actions having an influence time decay based on "the mean life time" which is the mean time for an action to propagate from one user to another, and can be calculated from training data. The joint influence probability changes with each new activation of a neighbor and needs to be recomputed after each step.

Discrete Time models are similar to the continuous model but simplify the assumptions. This model assumes the influence probability from a user to another user remains constant for a period of time after an action, after which it drops to zero.

The paper also proposes and interesting technique for predicting influence. But it assumes we have an action log, table of $(user, action, time)$ tuples where actions determine direct connections between neighbors in a graph, which is absent in our data. Instead, we will largely deal with pin timestamps to construct a graph-based model.

2.5 Pinterest Structure

Based on the work done by Gilbert *et al.* [1], repins are the primary form of interaction, as users don't generally directly interact, and follows do not happen nearly as often as repins. This allows us to measure diffusion in Pinterest using repins. Furthermore, we may be able to leverage textual information about users and boards to improve our predictions about pin propagation [6]. This would also help us fingerprint influential boards.

3 Dataset

The Pinterest dataset consists of a subset of US users who have food boards. The structure of the dataset is a dynamic bipartite graph of users-to-boards-to pins. This dataset contains over 10 million pinners, 12 million food boards, 736 million pins, and 48 million follows.

We load this dataset into four indexed tables in a Postgres database for easier and quicker manipulation and analysis. We set up the Postgres database on a dedicated server with 8 Intel Xeon cores, 32GB of RAM, and a 500GB SSD disk drive.

3.1 Data Characteristics

To characterize the boards we have, we categorized the boards and displayed the top 23 categories. Similar board names were aggregated together: e.g. all board names with 'fall', 'summer', ‘winter’, ‘spring’ in their name were categorized as seasonal_foods. Boards with generic names such as ‘Food’ were excluded.

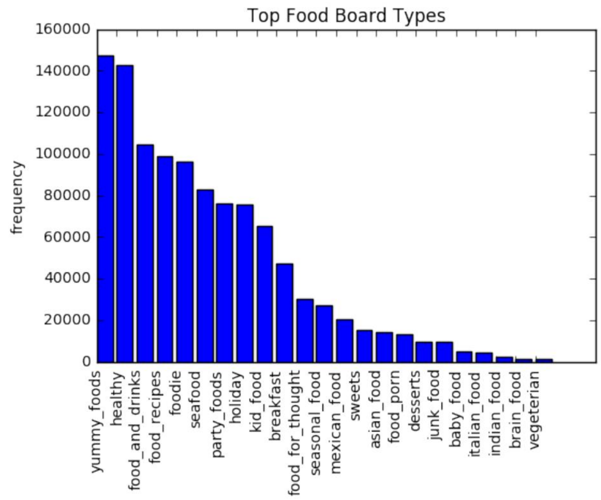

Figure 1: Types of food boards

To observe how often pins get repinned, we created distribution of the frequency of pins. This distribution follows a power law curve. The majority of pins (51%) never get repinned, and 85% receive fewer than 10 repins. This suggests that we want to focus our attention on pins on the upper end of this distribution because they better reflect influence propagation across Pinterest.

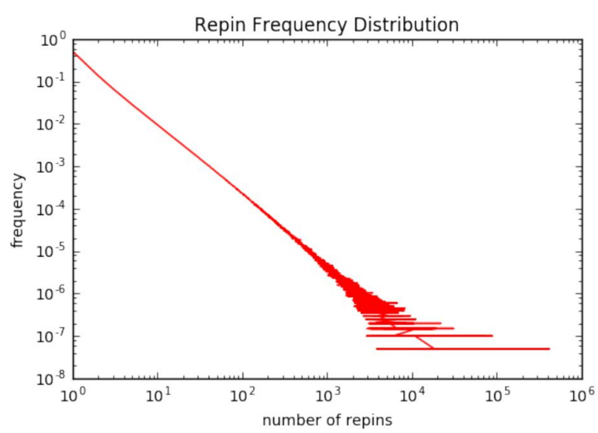

Figure 2: Distribution of number of repins

Furthermore, the distribution of the number of boards a user owns as well as the number of boards that a user follows in Figure 3 also follow a power law curve, which indicates that the majority of users do not follow many boards. Therefore, it may be in our best interest to reconstruct the network in which the nodes are boards, not users. In addition, the majority of boards do not have many followers, which hints that most repins come from boards that a user does not follow; building a graph based on followers would not provide a good model of influence on Pinterest.

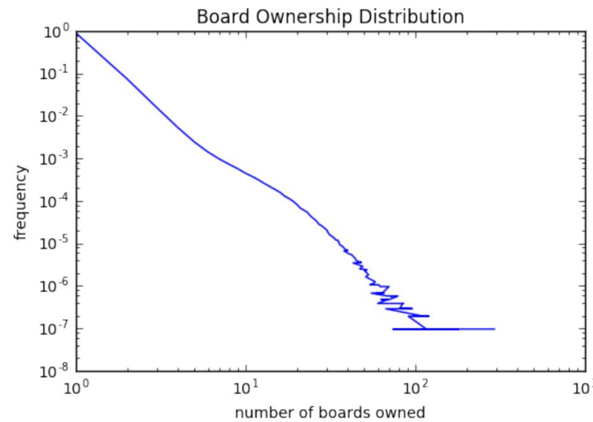

Figure 3: Distribution of board ownerships

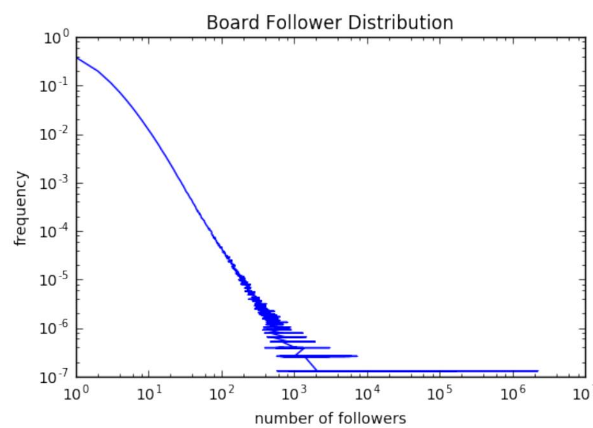

Figure 4: Distribution of board followers

4 Methodology

We formalize influence as the spread of pins in the Pinterest network through repins. For each pin, we only observe the time *when* it gets repinned and which board it was pinned on. We build upon the board latency model and the cascade model to understand the underlying network and explore the advantages of each model. We develop a temporal model on top of these graph models to represent influence as a dynamic process. We then analyze influence cascades on Pinterest according to these models. We utilize several graph statistics as well as random walks on these graphs to discover sources of influence.

4.1 Board Latency Model

We develop the concept of a board latency as when a board pins relative to other boards. For a given pin with a mean pin time of μ_p and standard deviation σ_p, we follow the technique of Gruhl *et al.* [4] by measuring the board latency for a board that repinned at time t as

$$b_l = \frac{t - \mu_p}{\sigma_p}$$

Negative values of board latency indicate that this board is ahead of the curve on this pin, whereas large positive values indicate that the board catches the end of a cascade. To evaluate a board's overall latency, we average it's latency across all of its pins.

In this model, an edge (u,v) is drawn with some probability p if board u comes before board v in some ordering, which is determined by creation times of a pin. If board u and board v share multiple pins, the edge has multiple chances of being created.

4.2 Cascade Model

Like the previous model, we interpret edges as connections between boards that frequently repin after another board.

We follow a similar model to Gomez *et al.* [3] where the probability of each edge is inversely proportional to the difference of their pin creation times. However, instead of using an exponential model, we use a power-law model. Formally, if we have the creation times t_u and t_v of nodes u and v, respectively, the probability $P(u,v)$ that edge (u,v) exists is given by

$$P(u,v) \propto \frac{1}{(t_v - t_u)^\alpha}$$

If $t_u \geq t_v$, we simply set $P(u,v) = 0$. We define a potential cascade c_i as a list of boards that contain pin i. This can be represented as a sorted list of times $[t_1, t_2, \ldots, t_n]$ when the pin was added to each board. We can view the cascade as a directed tree T where edges only go forward in time (of the pin).

The probability of observing cascade c_i in T is simply the product of all $P_{c_i}(u,v)$s for all possible (u,v) edges in T.

$$P(c_i|T) = \prod_{(u,v)\in T} P_{c_i}(u,v)$$

To find the probability of a cascade occurring in the entire network G, $P(c|G)$, we need to consider all possible trees T in which cascade could have occurred (all pin IDs whose repins follow a given cascade). This is given by

$$P(c|G) = \sum_T \prod_{(u,v)\in T} P_c(u,v)$$

Therefore, we want to find a global solution for all possible cascades occurring in the network and find the edges that maximizes

$$\hat{G} = \text{argmax}_{|G|\leq k} \prod_{c\in C} P(c|G)$$

where C is the set of all cascades in the network and k is the number of edges we want to infer. Intuitively, it will infer an edge (u,v) with high probability if board v consistently repins after board u within a small time window for multiple cascades.

To find a tractable approximate solution, we adapt the NetInf algorithm using the power-law model to infer the number of edges.

4.3 Temporal Analysis

We then select several interesting influential boards from the results of the models above, and perform a temporal analysis to observe how their properties change over time. Our algorithm starts by finding all pins of a given board, and for each pin's timestamp, we generate a "state of the world" sub-graph containing all of the boards that contain any of those pins at that point in time and all their pins posted on or before the timestamp. Each sub-graph can then be evaluated for various metrics such as betweenness centrality or cascade length. Because pins are not posted with a regular frequency, time is normalized by the respective board's pin posting sequence (first posted pin, second posted pin, etc.) instead of the absolute timestamp. Pins are sampled at even intervals from across the board's lifetime.

5 Results and Analysis

To observe cascades in this network, we first sample from our dataset, model information flow as a graph, and then analyze the resulting graphs.

The table of pins contains 95472102 distinct board ids, many of which have few (< 5) pins; we want to restrict our models to boards with a somewhat high number of pins (because finding an influential board with a low number of pins is more likely to be a fluke than a consistent result). Therefore, for each method, we sampled the top 10000 boards with the most pins. We then queried all of the pins that belong to these boards and reconstructed the underlying network.

5.1 Board Latency Sampling

For each pin involved in at least one of these boards, we chronologicallly order the boards that have this pin. Then, with some probability p, we draw an edge $b_1 \to b_2$, where b_1 comes before b_2 in the ordering.

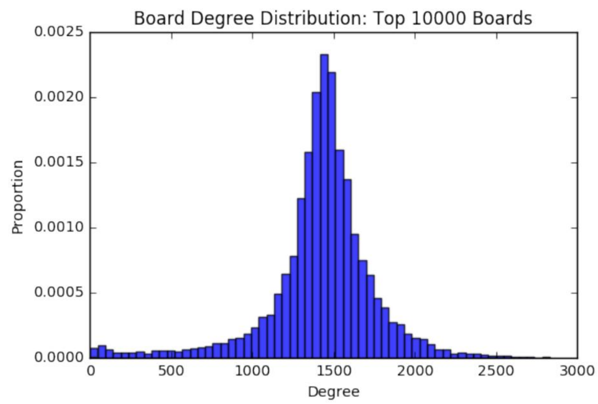

Figure 5: Degree Distribution, $p = 1$

We first set $p = 1$ to analyze the most connected graph structure possible. Having a high out-degree in this graph signifies that this board often pins before other boards, while having a high in-degree signifies that the board tends to repin from other boards. In this case, 9999 of the 10000 boards form one strongly connected component, with degree distribution roughly normal (Fig. 5). The out and in degree distributions look similar (normal), so this model suggests that repinning may not follow a preferential attachment method as we hypothesized.

As we reduce the edge probability p, we noticed that the SCC in this graph remained the same size. In fact, we had to drop below $p < 0.05$ to observe much reduction in the size of the largest SCC (Fig. 6). This suggests that the Pinterest network may have a bow-tie structure similar to what we observe in the web [7]. Because we are only looking at the boards with the most pins, these nodes would all fall in the large component at the center of the bowtie,

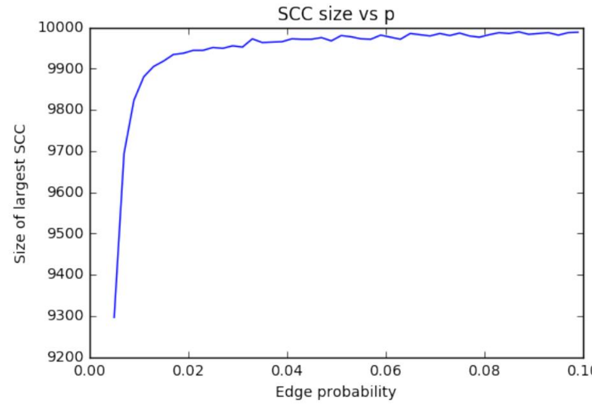

Figure 6: SCC Size

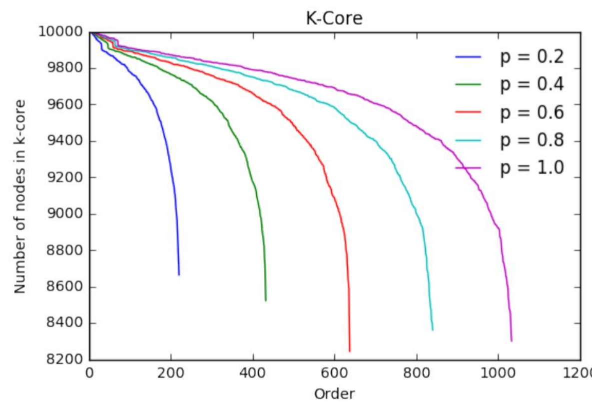

Figure 7: k-Core of Various Board Latency Graphs

which is why it takes such a small value of p to break this SCC. Furthermore, as seen in Figure 7, our k-core plots don't change shape as we change p, so we don't expect our graph structure to change significantly as we modify p.

Unlike the degree structure seen by Leskovec *et al.* [8], our graph is almost the exact opposite, as seen in in Figure 8. (Although the methods for constructing graphs are different in these two cases, we are just comparing the connectivity of boards/blogs here.) Following our model, this means that there are not many boards that are consistently leading other boards; therefore, we aren't likely to find many influential boards just by looking for boards that pin earlier than others; this model is therefore better suited for finding boards that amplify topics.

5.1.1 Optimizing Centrality

In this graph, betweenness centrality can be seen as a measure of how much the node controls the the flow of information. As seen in Figure 9, there is a

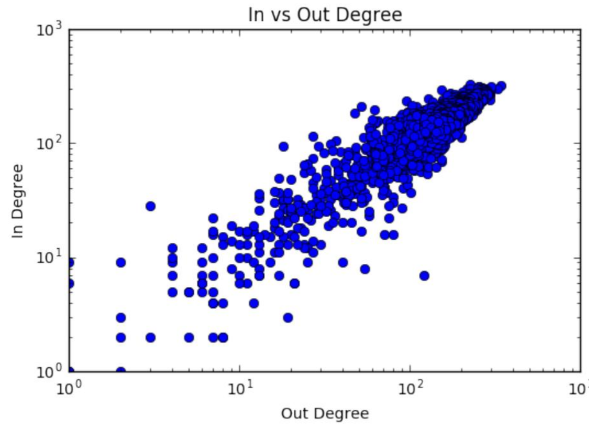

Figure 8: In vs. Out Degree, $p = 0.2$

weak correlation ($r^2 = 0.235$) between node degree and betweenness centrality. The nodes with low degree and high centrality are especially interesting because they indicate that the owner gained influence without many pins. For example, the point near the top left of the graph is a board titled "Christmas DIY, Games, Food & Decor" and has a total node degree of just 14, but a betweenness centrality of 1055, the 5th most in this graph.

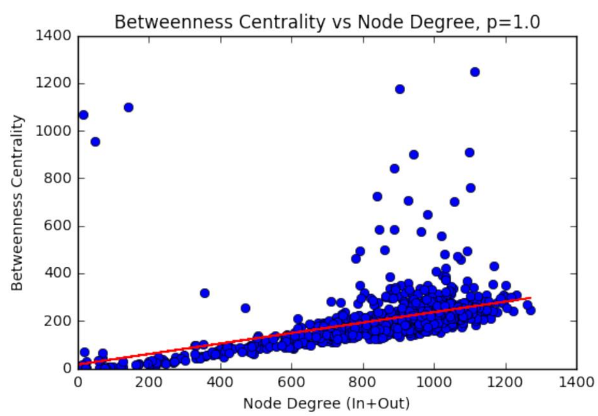

Figure 9: Betweenness Centrality, $p = 0.2$

This reinforces the observations of Ottoni *et al.* [1] that Pinterest constitutes several community structures, centered around topics. We categorize the neighbors of these boards, and in our qualitative observations, some boards (such as the Christmas board above) create bridges between multiple topics/communities, which affirms our findings in Section 3. To better observe this, we compared different instances of this board latency graph: we calculate betweenness centrality on a graph with $p = 0.2$, and compare with a graph with $p = 0.02$. We observe a weak negative correlation between betweenness centrality in the first graph and node clustering coefficient in the second graph. We therefore conclude that these nodes act as a weak 'glue' within the Pinterest network; they accelerate information cascades, but aren't crucial. However, these boards still have a high level of influence because they can select what other pinners are likely to see and which topics will trend sooner.

5.1.2 Centrality and Latency

When we analyze board latency versus centrality, we observe that boards ahead of the curve and behind the curve do not lead to high centrality. However, the boards with the highest centrality generally lie within one standard deviation of the mean in board latency. As we hypothesized above, the boards with high centrality are often not the first to pin; rather, they catch the middle of many trends and amplify them. This is consistent with our earlier observations, where central boards relate multiple topics, so they would introduce a pin to a community that hasn't yet seen that pin. In a sense, this can be thought of as 'information arbitrage': these boards gain influence by taking existing ideas from one community and introducing them as new ideas in another community. This evidence is consistent with our belief that one effective technique of maximizing influence on Pinterest is bridging multiple communities/topics.

5.1.3 Random Walks

While our plot of node degree earlier indicated that there may not be boards that are consistently earlier than other boards at pinning, we would still like to know which boards to observe to know the volume of pins that flow through the network. In an effort to find the 'source' of cascades, we could search for source nodes in the graph. However, this would not be a good approach as boards with just one pin that got repinned would appear as sources. Furthermore, any cycle would rule out all nodes in that cycle. Therefore, we take random walks through this graph, stopping with a probability $p = 0.05$ and teleporting with probability 0.2. We record how often the random walk stops at each node. This distribution is roughly normal and highly correlated with out-degree (Fig. 10).

The strong correlation with node degree indicates that boards that pin more often are likely to be sources, and that boards rarely deviate from this (so one wishing to gain influence on Pinterest is unlikely to be able to introduce new pins early and

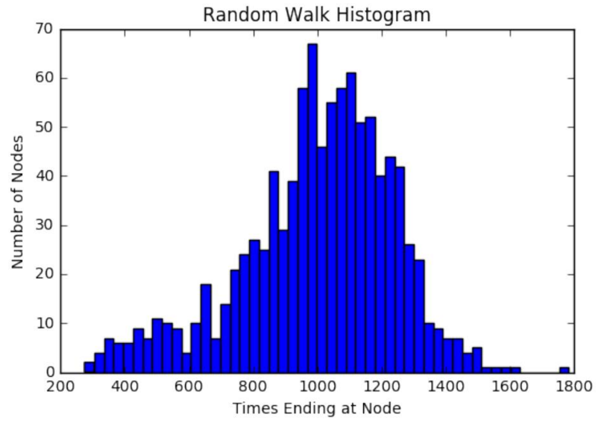

Figure 10: Random Walk Distribution

initiate large cascades through the network). This is also consistent with the findings of Bakshy *et al.* [2], which is logical given the nature of 'repinning' versus 'retweeting'. Furthermore, looking into the boards where the random walk ended most often, the top board is named "Food & Drink" and the next dozen are named variations of Food, which would represent the center of our network in terms of topic. These boards also do repin from one another, with the k-core holding together even at fairly high orders and low probabilities, as seen in Figure 7.

Therefore, this model indicates that influential boards either initiate cascades with a high volume of pins, and amplify cascades by bridging communities. However, while this approach scales with the amount of data we have, one major shortcoming is that it is unable take time differences into account, which we address with our model in the following section.

5.2 NetInf

Using the cascade model, for each unique pin in the set of pins among the top 10000 boards, we grouped all boards for each unique pin i. This formed the cascade c_i, where creation times were sorted in ascending order. To remove cascades that we consider to be negligible, we remove all of the pin cascades that have have a length of one.

For this set of boards, we observed 581,253 different cascades. We performed the NETINF algorithm with parameter $\alpha = 0.0001$ and run for 3000 iterations, which inferred $\sim$ 3700 edges.

The average number of cascades that an edge was part of was 30, which is higher than the average in [3], which had an average of 7, which gives us higher confidence in our edges. The number of edges per cascade follows a power law distribution.

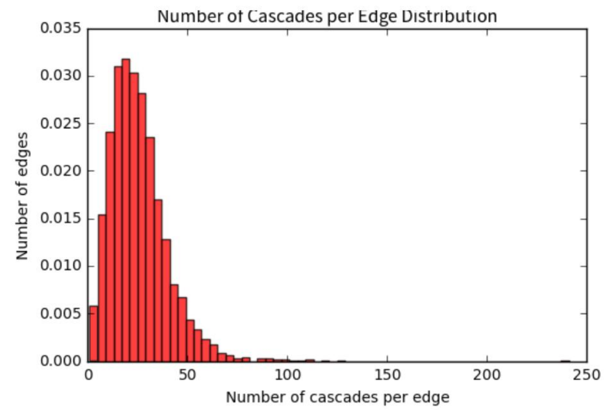

Figure 11: Distribution of Cascades per Edge

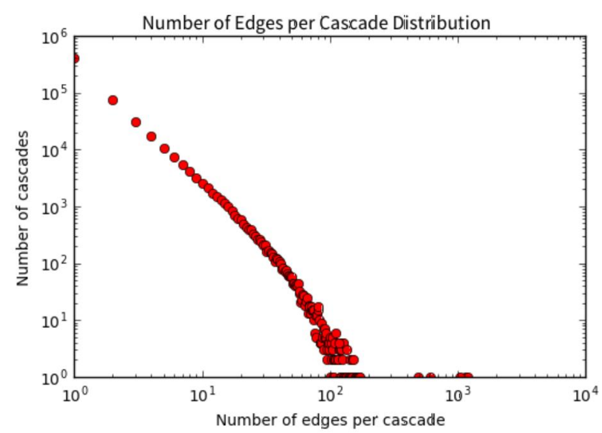

Figure 12: Distribution of Cascade Sizes

We notice several differences from the previous board latency model. First, the degree distribution of the resulting network closely follows the Preferential Attachment model, whereas the board latency model is closer to those of random graphs.

Furthermore, there is not a strong positive correlation between the in- and out-degree distribution (Fig, 13). This plot is similar to the observations of Leskovec *et al.* [8]; therefore, NetInf indicates the Pinterest network may resemble a webblog network. As opposed to the board latency model, our figure shows that there are many boards who have small in-degrees with large out degrees and boards with large in-degrees with small out-degrees. This suggests that many boards that are sources in the network can be categorized as sources of influences.

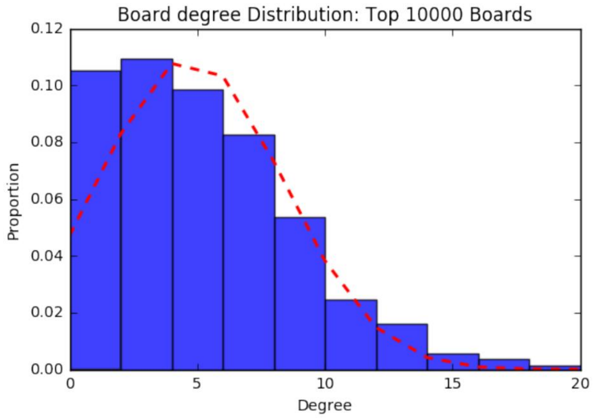

Figure 13: Degree Distribution of Network inferred by NetInf

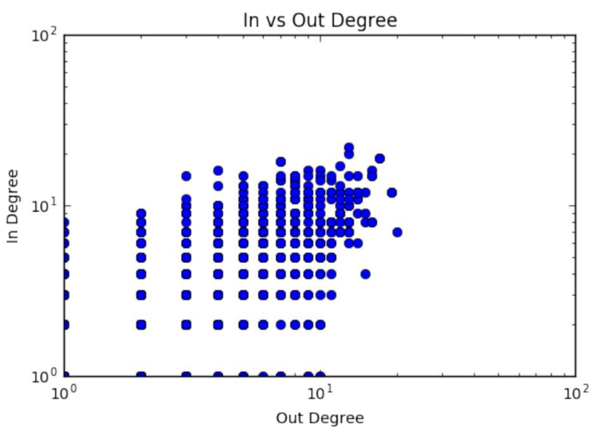

Figure 14: In vs. Out Degree (NetInf)

5.2.1 Sources of Influence

From our resulting graph, we analyzed the characteristics of nodes with 0 in-degree and whether they corresponded to initiators of influence. These source nodes do not correspond to the top boards that have the most pins.

For all boards that were identified as source nodes, we found the number of cascades where the first pin in each cascade belongs to a source node. Because a source of a cascade may not be the very first pinner, we also relax our definition of 'source' of cascades to boards that pinned within the first 5 repins.

We observed that although boards that were source nodes make up only 0.047 of all boards, it has a much higher percentage of appearing within the first 5 repins. Furthermore, the average cascade size for cascades in which the source nodes appears within the first 5 repins is much higher than the overall average cascade size. This suggests our model accurately identifies sources of influence, which we often recognize as those whose influence lasts for a long time. However, it is interesting to note that the average out degree of these sources nodes is much lower than on average. This result differs from that of the board latency model, which suggested boards with high degree are often sources.

	Source	All
Num Boards	470	10000
First pin of cascade	34158	581K
Pct. of Cascades (first pin)	0.08	1.0
First 5 pins of cascade	104K	581K
Pct. of Cascades (5 pins)	0.19	1.0
Avg. Cascade Size	2.79	2.29
Avg. Path Size	3.90	3.02
Avg. Out Deg	1.80	4.78

Table 1: Cascade Comparisons between Source vs. All Boards

We realize this is a consequence of how the edges were formed in each model. In the cascade model, it conservatively chooses edge where a board v repeatedly repins after a board u. Despite the differences in how these edges were formed, the board types of these sources were consistent, where a majority of the boards were named generic names such as "Food" and "Food & Drink".

Following our analysis for the board latency model above, we also attempt to find 'sources' of cascades using random walks. We observe a skewed right distribution (Fig. 15), which is more in line with what we expect compared to the distribution from the board latency model. The boards with high scores from random walks were similar to the source nodes in the graph. We analyze a sample of these boards in the section.

5.3 Influence over Time

As a final step, we explored the temporal nature of influence and how it changes over time. For both the Board Latency and the Cascade models of influence, we took samples of a representative influential node along with a non-influential node with a similar number of pins and compared their properties to determine whether there was a measurable difference.

As stated previously, betweenness centrality is a useful measure of influence in the board latency

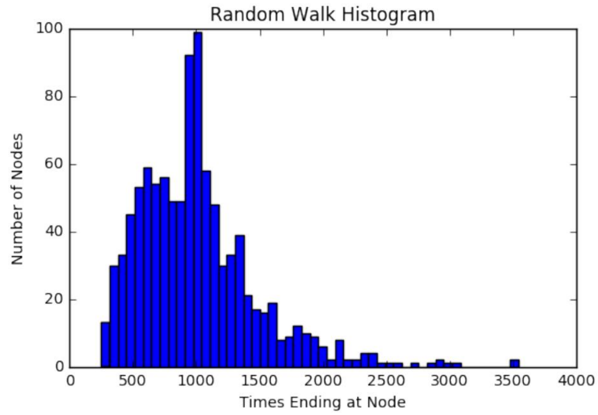

Figure 15: Random Walk Distribution, NetInf

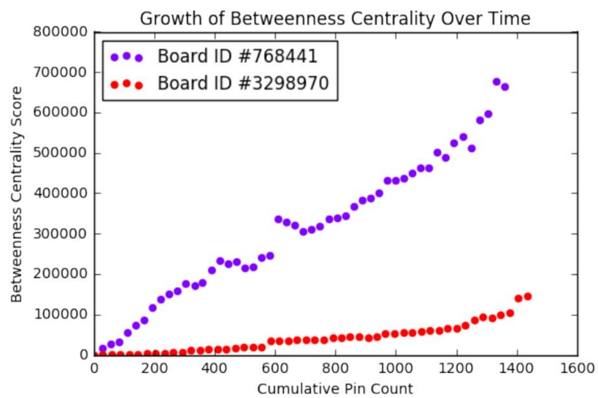

Figure 16: Temporal analysis of the board latency model shows how betweenness centrality grows much quicker over time in influential boards

model, as it shows the importance of a board as an information arbiter. A high betweenness centrality can be used as a proxy for the influential status of a board in between communities. In particular, Fig 16 reveals two insights that would be valuable to marketers. First, the higher marginal gain in betweenness centrality for influential boards is apparent from the beginning of the board's existence. The outcomes of this clean trend are that (assuming a consistent pinning pattern) it is possible to forecast a board's future influence with the board latency model, and also classify it as being influential or not. Second, the chart shows how effective the "community bridging" approach can be in gaining influence on Pinterest (purple plot). The influential board's influence grows an order of magnitude faster than its non-influential counterpart. To give a concrete example, after 50 pins the influential board enjoys a betweenness centrality score equal to that of an "average" board after 500 pins!

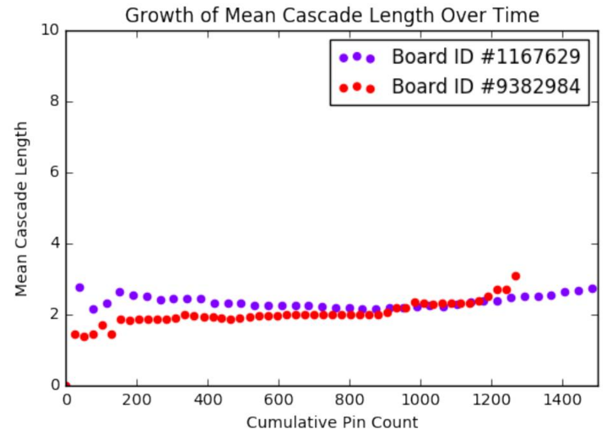

Figure 17: Temporal analysis of the cascades model shows that influence via mean cascade length changes primarily as a function of number of pins

Mean cascade length is another way to measure influence. A growing mean cascade length over time may signify that the board is closer to the heart of the community, because the average cascade length would be closer to 0 as a board approaches the fringes of the network. It can also signify that the board is a popular "content creator", surfacing unique content that then goes viral, and gets shared many degrees out from the original source.

The examples in Fig 17 containing two source nodes show that becoming influential as a "content curator" on Pinterest is not trivial. With over 10 million users, and close to 750 million pins in the food category alone, finding or creating novel content that will be repinned is more a matter of luck than skill. Aspiring Pinterest marketers would be well heeded to not waste time trying to create the perfect content to pin onto their board, but rather take the approach of re-pinning existing content with high volume, or finding content to pin at the bridge of two communities. More detailed data on re-pins would be needed to make a definitive conclusion.

6 Conclusion

The board latency model provided a good model of how influence flows on Pinterest. We observed that Pinterest contains community structures centered around topics and that many boards with high betweenness centrality serve as bridges between these topics. This leads us to the conclusion that one method of becoming influential is through 'information arbitrage'; by replicating pins from one community to another, a board can gain influence without needing to initiate cascades.

We also attempted to find sources of cascades under this model using random walks, which suggested that these sources have high out-degree. This surfaced the shortcoming of our first model because it couldn't take time differences into account. However, these boards provide insight into the state of the Pinterest network because they include many pins involved in cascades through the network.

In contrast, the cascade model determined sources of influences by directly identifying the sources of the graph. Interestingly, these sources did not have high out degrees, which indicates that sources influence other boards which amplify their influence. This is consistent with our conclusion from the board latency model that many boards with high in- and out-degree serve as bridges of influence. Furthermore, this matches our conclusion from our k-core analysis that indicated boards that pin early are not necessarily the most influential; boards that repin later with a high volume of pins may do better in terms of influential gain. Therefore, sources may initiate cascades, but require high centrality boards to amplify their content to render the cascade effective throughout hte network.

We conclude that marketers may be interested in identifying and creating boards which propagate cascades and connect multiple niche communities, such as a board focusing on Christmas food crafts. Otherwise, the typical approach would be to create a board with a high volume of pins.

7 Future Work

Having additional board/pin data (such as international users) and re-pin sources, would increase the accuracy of our analysis. Furthermore, because have evidence that there are strong communities revolving around topics in the network, analyzing data beyond just food boards would provide more insight into these communities. We have determined two main methods of influence: initiating cascades and amplifying them. An interesting follow-up would be how these two types of users could collude to help one another gain influence on Pinterest.

References

[1] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. "i need to try this"?: A statistical overview of pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2427–2436, New York, NY, USA, 2013. ACM.

[2] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[3] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *CoRR*, abs/1006.0234, 2010.

[4] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[5] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.

[6] Raphael Ottoni, Joo Paulo Pesce, Diego Las Casas, Geraldo Franciscani Jr., Wagner Meira Jr., Ponnurangam Kumaraguru, and Virgilio Almeida. Ladies first: Analyzing gender roles and behaviors in pinterest, 2013.

[7] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(16):309 – 320, 2000.

[8] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining (SDM07)*, 2007.

8 Individual Contributions

Michael: Wrote the board latency model and random walks, along with corresponding analysis
Sam: Wrote the NetInf model and analysis
Lucas: Setup Postgres server, wrote the temporal model with analysis