Analyzing WeChat Diffusion Cascade: Pattern Discovery and Prediction

Ruilin Lv, Chengxi Zang, Wai Kin Victor Chan and Wenwu Zhu

July 18, 2019

Analyzing WeChat Diffusion Cascade:
Pattern Discovery and Prediction

Ruilin Lv[1], Chengxi Zang[2], Wai Kin Victor Chan[*1], Wenwu Zhu[1]

[1] Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
Shenzhen, 518071, P.R.C.

[2]Department of Computer Science and Technology
Tsinghua University
Beijing, 100084, P.R.C.

Abstract

WeChat social network is one of the most popular social platforms in China, providing not only communication services but also enabling a number of service innovations. Understanding how information diffuses in an online social network such as WeChat is critical to the design and evaluation of existing or new services. This paper studies the diffusion pattern and predictability of WeChat cascade. We propose an analysis framework for WeChat cascade based on the characteristics of cross-scenario diffusion. By analyzing a real WeChat dataset, we reveal some typical diffusion patterns. We also obtain good prediction performance.

Key words: social network, WeChat, information diffusion, prediction

1. Introduction

The boundary of service systems is disappearing with the emerging of various kinds of new services thanks to the advancement in information technology such as IoT, social networks, and so forth [1]. The increasing popularity of social network platforms such as Facebook and WeChat not only facilitates communication between people but also creates new service innovations. For example, WeChat is providing many mini-apps that provide all kinds of services to users, including payment, finance, food, entertainment, travel, education, and so forth.

In China, WeChat is probably the most popular platform for online social interactions currently. It has more than 1 billion Monthly Active Users. Almost all Chinese net citizens use WeChat to communicate with friends and obtain multimedia information. WeChat offers multiple ways for information diffusion, including Chat and Moments, which inherently leads to high complexity of information diffusion. Therefore, understanding the diffusion patterns of WeChat can help improve the design of online services.

Using real WeChat diffusion dataset, this paper analyses the diffusion process of WeChat articles by incorporating diffusion cascade and user behavior record. First, we obtain the diffusion pattern of WeChat articles. Then, we use machine learning algorithm to predict future evolution of WeChat diffusion. Finally, using visualization, we perform case studies on cascades in different content topics.

There are existing studies on diffusion mechanism of social platforms like Facebook, Twitter, and Weibo [7, 10, 11, 12, 13, 14]. However, the study of diffusion mechanism in WeChat network is still a relatively new area. This paper tries to enrich this research area by analyzing the WeChat data. It was found that WeChat network is rich in diffusion patterns. WeChat exhibits a cross-scenario diffusion characteristic that makes its structural pattern of diffusion cascade full of complexity. This paper uses advanced machine learning algorithms to predict WeChat diffusion process. It was found that content topics might be an important factor that influences the evolution of the diffusion structure. Results of this paper add values to research on analyzing information diffusion and communication behaviour in WeChat network. The results may be used to better design social networks to achieve beneficial goals such as blocking spreading of rumors, promoting positive emotions, and even facilitating new service innovations.

[*] Corresponding author: Tel.: (86) 3688-1023; E-mail: chanw@sz.tsinghua.edu.cn

2. RELATED WORKS

This section reviews some important work in social network analysis and information diffusion process. In social network analysis, a well-known phenomenon: “Six-degree-of-Separation” was proposed and experimentally verified [2]. Under this phenomenon, a person could reach any other one in the world along the relationship chain within 6 people on average. The Small-world Theory [3] introduced a model via re-wiring to capture characteristics of real world networks, such as local clustering and short distance.

For network structure, Barabasi and Albert proposed the scale-free network model in which degree distribution follows a power law. This model relies on the preferential attachment and growth process that generate a network with a few “hub” users, which also represents many real world networks [4]. Community Detection is another hot research topic that studies heterogeneous communities [5].

The process of information diffusion is one of the key research questions in social network analysis. In order to explain how information diffuses across social networks, traditional explanatory model can be used. One example is the Epidemic Model. In this model, each user can receive and distribute diseases/information, which is governed by a certain probabilistic infection mechanism [6]. Another explanatory model is the Social Influence model, which evaluates and maximizes the influence of individual (or community) in a social network [7,8].

One major goal of the study of information diffusion is to predict the diffusion process in advance. Predictive models concentrate on improving the predictability of information cascade. Along this line of research is the independent cascade model, which regards the diffusion process as a tree structure and consider users who contribute to information spreading. Ref. [9] claimed that the cascade predictability was determined by “skill” and “luck”, and discussed possible sources of prediction error. Ref. [10] examined the prediction error when adding different categories of features and explored the limitation of prediction accuracy. Ref. [11] considered the diffusion process based on truth and rumor, and compared the cascade structural pattern by using a Twitter dataset. Ref. [12] explicitly defined the cascade prediction problem to avoid the side effect of sample imbalance. Ref. [13] proposed the metric of structural diversity and distinguished viral cascade from non-viral ones. Ref. [14] presented other alternative structural metrics, such as the number of connected components within a user’s friends network.

There are also other predictive models. For instance, Linear Threshold Model and Game Theory Model. The former model focuses on the study of the triggering threshold of diffusion [15], while the latter one concentrates on equilibrium analyses of among information spreaders [16].

In summary, existing works have built a basic research framework for the analysis and prediction of diffusion cascade in social networks. This paper contributes to this area by providing an in-depth study into the diffusion process of WeChat network.

3. BASIS OF WECHAT DIFFUSION RESEARCH

3.1. CROSS-SCENARIO DIFFUSION CHARACTERISTIC

WeChat has three diffusion scenarios, which make the diffusion process different from those in other social platforms (such as Weibo, Facebook, and Twitter). These three diffusion scenarios are: initial release scenario, chat scenario, and moments scenario.

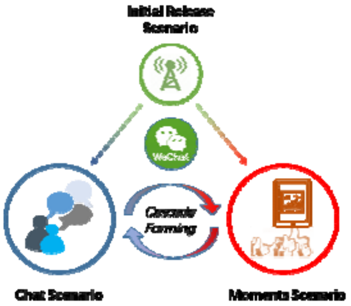
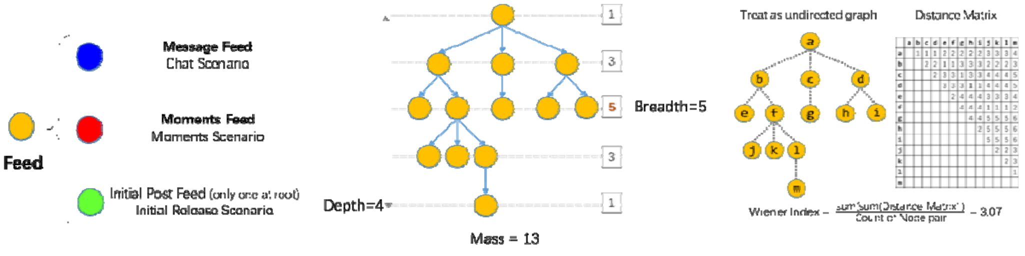


Figure 1: Cross-scenario diffusion characteristic of WeChat.

WeChat diffusion process cascades step by step as follows: An article is initially published by Official Account. The link of this article first appears in the Initial Release Scenario. Then subscribers are able to share this link to both Chat scenario and Moments scenario, thus cascading the diffusion of the article.

3.2. *Measuring WeChat Cascade Structure*

WeChat cascade is composed of "feeds". Each feed is a link to an article. The feeds can appear in any scenarios. In order to evaluate the influence and structural complexity of WeChat cascade, four basic metrics are employed, including mass, breadth, depth, and wiener index.



	a	b	c	d	e	f	g	h	i	j	k	l	m
a		1	1	1	2	2	2	2	2	3	3	3	4
b			2	2	1	1	3	3	3	2	2	2	3
c				2	3	3	1	3	3	4	4	4	5
d					3	3	3	1	1	4	4	4	5
e						2	4	4	4	3	3	3	4
f							4	4	4	1	1	1	2
g								4	4	5	5	5	6
h									2	5	5	5	6
i										5	5	5	6
j											2	2	3
k												2	3
l													1
m													

(a) Three types of feed (b) Metric of Mass, Depth and Breadth (c) Metric of Wiener Index

Figure 2: Component feed and metrics of WeChat cascade.

Mass is defined as the number of feeds in cascade. Mass value reflects the level of overall diffusion coverage. Therefore, the order of magnitude of mass is a main measure for the diffusion performance.

Breadth is defined as the maximum number of feeds that single cascade layer holds. Breadth value reflects the horizontal size of cascade. Larger breadth could imply the characteristic of "broadcast diffusion". The order of magnitude is also a main measure.

Depth is defined as the maximum distance between leaf nodes to root node. Depth value equals the number of layers, indicating the penetrability of information diffusion. Larger depth value suggests that the cascade have at least a few long chains that enrich its diffusion structure.

Wiener index is defined as the average distance of all node pairs. We need to obtain cascade wiener index by treating the cascade in undirected view, then construct the distance matrix and apply the division formula. Wiener index reflects the overall structural complexity and provides a quantitative measure of the structural pattern of the diffusion.

3.3. *Dataset Description*

Our dataset includes 7504 independent diffusion cascades. Each cascade corresponds to one article randomly sampled from overall WeChat Official Account published content. The corresponding articles were released between March 1st and March 7th in 2016.

The dataset is composed of five data fields, including sharing cascade record, article reading record, user relationship record, user attribute record, Official Account subscribers record.

4. Discovering Diffusion Pattern in WeChat

4.1. *Distribution of Structural Metrics on Final Cascade*

We first examine the frequency distribution of WeChat cascade structural metrics. From the complementary cumulative distribution function curve in logarithmic coordinate system, it is observed that the distribution of these structural metrics are extremely uneven. The major cascades are in small mass and breadth, with low depth and Wiener Index. However, there exist some extremely huge cascades in terms of coverage size and structural complexity. There are nearly 1% of cascades that could be called "enormous" as their mass exceeded 10,000. This situation is similar to the distribution of breadth value. For metrics of depth and Wiener Index, the largest observed values are around 33 and 20, respectively, confirming the existence of a highly structural and complicated cascade in WeChat diffusion.

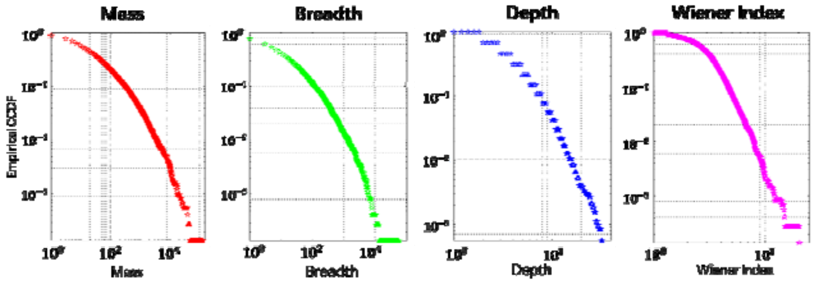


Figure 3: Distribution of final cascade metrics value.

4.2. *Correlation Pattern of Structural Metrics*

Next, we analyze the correlation between mass and other structural metrics. The heat maps in Figure 4 reveal several findings. First, breadth and mass values are positively correlated. This means that if a WeChat cascade is large in mass, it has very high probability to hold high breadth value at a similar magnitude. Mass is also found to have positive correlation with depth and Wiener Index. The heat map presents a diverging trend especially when mass is large. Cascades tend to hold lower depth and smaller Wiener Index when their mass value is below medium. This phenomenon may be explained by using the uncertainty theory. Once a WeChat cascade grows larger in mass, it suffers from more uncertainty in structure evolution. The current cascade could either stay in simple structure or transfer to complex structure.

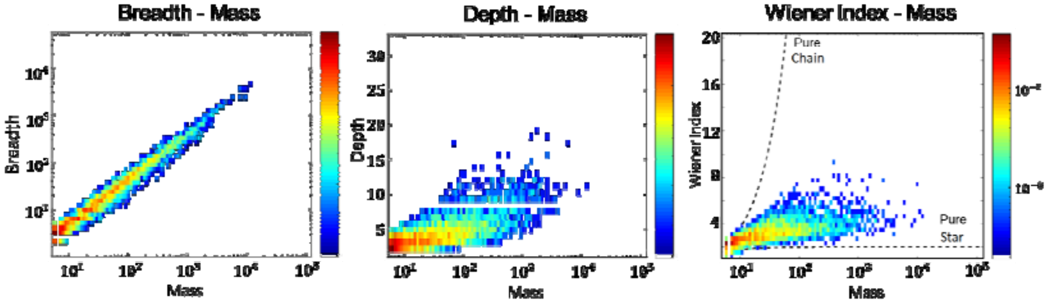


Figure 4: Joint distribution of structural metrics to mass, plotted in heat map.

In order to distinguish whether a WeChat cascade is structurally simple or complicated, we define the structural patterns as "star-like" and "chain-like". The numerical indicator is Wiener Index. Our definition starts from the boundary condition: if the mass is *n*, the Wiener Index value of "pure star" is *(2-2/n)*, and *((n+1)/3)* for "pure chain" structure. From a mathematical point of view, "pure chain" is structurally more complicated than "pure star" at any mass value *n*. From the empirical dataset, the structural complexity of most WeChat cascades are somewhere in between "pure star" and "pure chain", except for a few extremely small ones. The distribution area is restricted between the asymptote of "pure star" and "pure chain" as shown in the right heat map. It is observed that, the major structural pattern of WeChat cascades is closer to star pattern, especially for large cascades. In addition, with the increasing uncertainty in diffusion process when cascade mass grows larger, the element of chain is gradually added into the cascade structure, leading to a larger variance of Wiener Index.

5. Predicting Cascade Structural Evolution

5.1. *Feature Extraction*

Our empirical dataset provides information about cascade diffusion and related user behavior. After preprocessing and feature engineering, six categories of useful features are extracted for each cascade. They would serve as the input for prediction in the next subsections. All these features are captured in cascade midway at mass of *k*.

First category is *Content Feature*. This category includes three features as follows: article content length (in byte count), subscriber amount of Official Account, post time indicator (to distinguish A.M. post or P.M. post).

Second category is *Cascade Structural Feature (at mass of k)*. This category includes three features as follows: cascade depth, cascade breadth, cascade Wiener Index.

Third category is *User Relationship Graph Feature*. As the relation graph is very sparse, two extracted features are considered: number of connected components, and fraction of users belonging to the largest connected component

Fourth category is *Forwarding User Attribute Feature*. Features in this category are only about those users who have contributed to the cascade. This feature category provides user identity record, such as average age, sex ratio, average educational level, average account activation years, and average number of friends. In addition, this category also consists of features that reflect a user's activity level on WeChat in the recent month (i.e., March 2016), including average amount of chat messages sent and received by this user, and average amount of Moments post and view by this user.

Fifth category is *Reading User Behavior Feature*. As some users may not share an article with others after reading it, this feature category evaluates the relationship between reading and forwarding: reading frequency of all reading users, reading frequency of forwarding users, ratio of users who have read the article more than once, ratio of reading platform (mobile/PC), user conversion rate (from reading to forwarding).

Sixth category is *Temporal Feature*. This feature category embodies the effect of dynamics, including time elapsed (since original post), average time gap between each two consecutive feeds (from the first feed to the k/2 feed, and from the k/2 feed to the k feed), and the variance of time gap.

5.2. *PROBLEM DEFINITION*

The objective of our prediction task is to achieve a high prediction accuracy and stability of the final diffusion coverage and structural complexity when a cascade is at its beginning stage. Specifically, the prediction problem is defined as a binary classification task: for a given cascade, predict whether the final metric value (mass, depth, breadth, Wiener Index) of this cascade would exceed the median value comparing with its peers that also currently hold a mass of *k*. Using the median value to distinguish positive and negative samples can balance the sample ratio, thus reducing the risk of sample imbalance.

5.3. *PREDICTION PERFORMANCE OF MACHINE LEARNING MODELS*

To perform the binary classification task, several machine learning algorithms are tested, including Logistic Regression, Naïve Bayes, Support Vector Machine, and Decision Tree. Furthermore, in order to improve the model performance, ensemble learning methods like bagging and boosting are applied to Decision Tree, obtaining the Adaboost and Random Forest model. The observation size is *k=100* and F1-score is used to evaluate the classification performance.

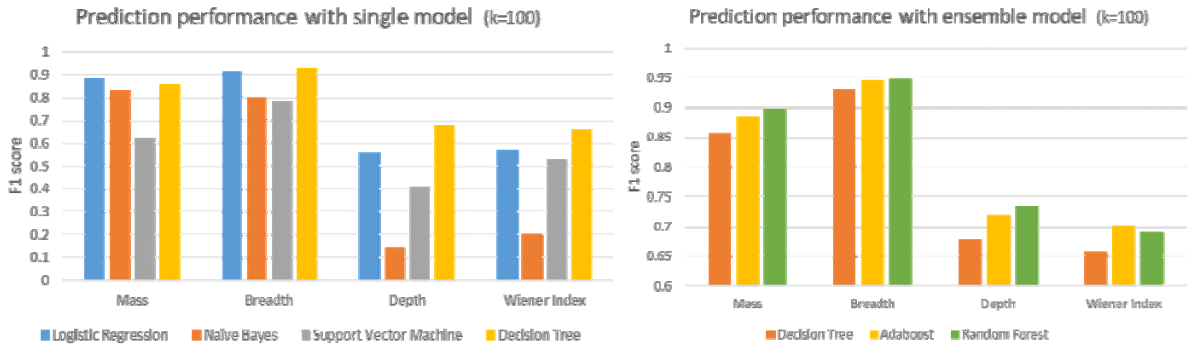


Figure 5: Prediction accuracy with different machine learning algorithms, in single model and ensemble model.

Figure 5 gives the prediction performance of single and ensemble models. For single models, Decision Tree and Logistic Regression generally perform better than the other two models. As for the predictability of the four metrics, breadth and mass are obviously more predictable than depth and Wiener Index. Decision Tree is generally a reliable model, which gives F1-scores of 0.85, 0.92, 0.68, 0.65 for the four metrics, respectively.

The prediction performance of ensemble models suggests that, both Random Forest and Adaboost performs better than the single Decision Tree model, while the performance of Random Forest is slightly better. For instance, in the depth prediction task, Random Forest model could improve F1-score from 0.67 to 0.73, compared to the single Decision Tree model. In addition, based on bagging method, Random Forest model is efficient to deploy in application with the help of parallel computing. Therefore, Random Forest model is used in the following prediction task.

5.4. Prediction Performance With Feature Set And Observation Size

In real application, observation size *k* and selected feature set are two key factors affecting system performance. It is desired to obtain a certain level of prediction accuracy by consuming the least computational resources and detecting abnormal diffusions as early as possible. To obtain a certain level of prediction accuracy, smaller size of feature set and smaller observation size *k* are preferred. Therefore, the prediction performance is tested under different observation sizes and different single feature categories.

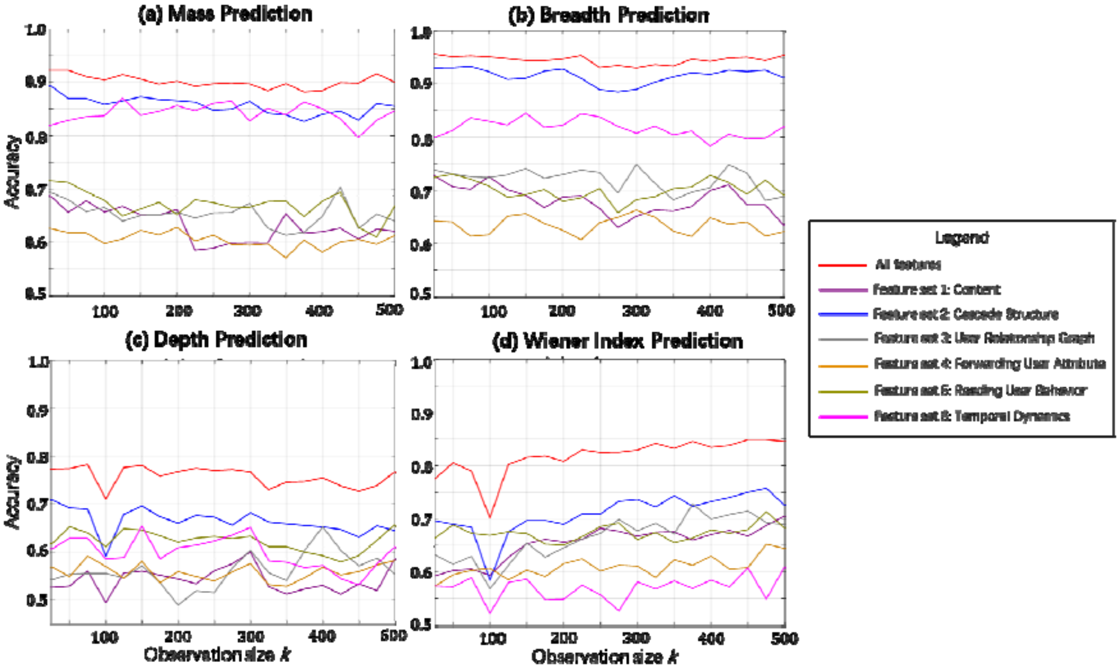


Figure 6: Evolution curve of prediction accuracy with different input feature set and observation size.

In predicting mass and breadth, it is found that, the complete feature set performs the best, with a prediction accuracy around or higher than 0.9 at all observations. When single feature set is considered, structural features and temporal features give the top two performances. This indicates that the current structure and dynamics are intrinsically reflective to future coverage of cascade. For other feature categories, prediction accuracy varies from 0.6 to 0.75, due to large fluctuation.

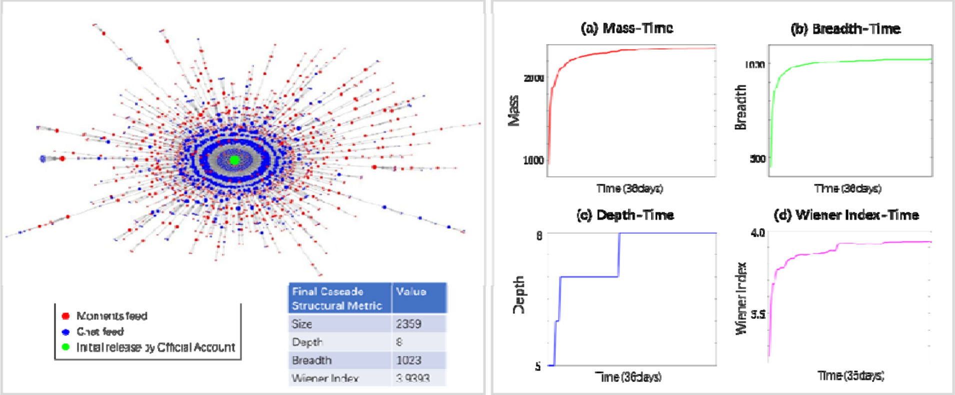
For the prediction of depth and Wiener Index, the complete feature set performs the best, with a prediction accuracy around 0.8. The performance of single feature sets is generally poorer and has no essential difference among each other. For single feature categories, prediction accuracy ranges from 0.5 to 0.7, also suffering from large fluctuation as the observation size increases. In addition, an apparent trough at k=100 is detected in almost all curves. This observation implies that k=100 might be a key demarcation point of predictability for cascade structural complexity. Once cascade mass grows larger than 100, the structural predictability goes back to its early stage.

In summary, first, mass and breadth are more predictable than depth and Wiener Index. Second, the complete feature set leads to the best prediction performance, while the structural and dynamical features set seem to contribute a lot. Third, observation size *k* has little impact to prediction performance, confirming the robustness of our prediction model.

6. Case Study of WeChat Cascade

6.1. Case one: “Star-like” cascade

Case one examines the typical “star-like” pattern. The article content is public instruction about how to choose best insurance to buy. The content topic could be categorized as “Life Guidance”.



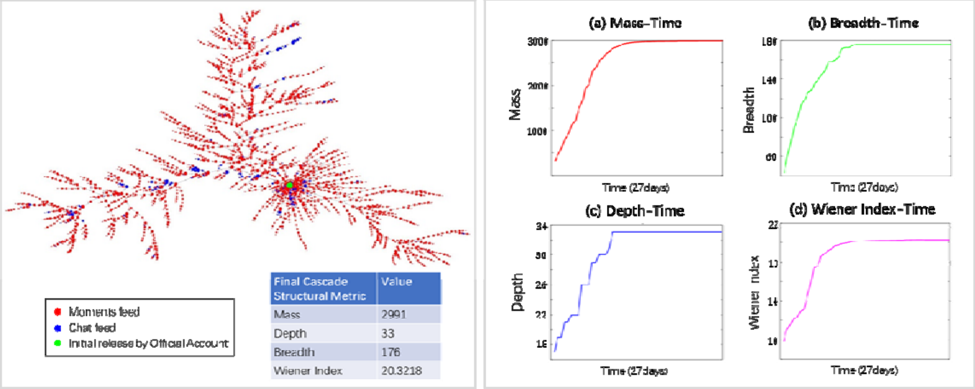
Final Cascade Structural Metric	Value
Size	2359
Depth	8
Breadth	1023
Wiener Index	3.9393

Figure 7: “Star-like” case with visualization of structure and dynamics

In this case, the cascade structure presents a “star” like shape. From the visualized topology in Figure 7, it is clear that the cascade is driven by two scenarios in balance, as the message feeds and Moments feeds account for similar proportion. Based on the statistics of final structural value, this cascade has medium mass value, low depth value, high relative breadth (the widest layer holding nearly half of all feeds), and low Wiener Index. The structural complexity is relatively low. From the dynamical evolution curve on timeline, the initial bursts in terms of all metrics are obvious. This observation indicates that, the major diffusion process has quickly come to the completed stage within only a few days after initial release. One explanation is that this topic of content could have drawn attention from all adults with different backgrounds. Therefore, the diffusion path is uniformly distributed on the underlying social network, presenting “star-like” pattern in final structure.

6.2. Case two: “Chain-like” cascade

Case two examines the typical “chain-like” pattern. The article content is commercial advertisement driven by induced forwarding in title. The content topic could be categorized as “Commercial Marketing”.



Final Cascade Structural Metric	Value
Mass	2991
Depth	33
Breadth	176
Wiener Index	20.3218

Figure 8: “Chain-like” case with visualization of structure and dynamics

In this case, the cascade structure presents a “chain” like shape with several hierarchical branches as shown in Figure 8. The cascade is dominated by the Moments scenario diffusion. Based on the statistics of final structural value, this cascade has medium mass, very high depth, quite low breadth value, and extremely large Wiener Index. The complexity of overall structure is rather high. The dynamical evolution curve presents a smooth growth process in terms of each metric value. The cascade consistently grew deeper and generated more branches along the timeline. This content is designed to attract specific people to buy certain goods, by using the approach of collecting “likes” in each user’s Moments post. Therefore, it is not surprising that this cascade is Moments-scenario driven and diffused along the relatively determined path.

7. Conclusions

This paper built an analyzing system framework to discover patterns, test predictability and perform case study on WeChat information diffusion cascades. Several typical diffusion patterns emerged in terms of cascade structure, user behavior, and content topic. This finding indicated that the diffusion process on WeChat network has a huge complexity. The prediction performance obtained by advanced machine learning algorithms was satisfactory, which revealed the predictability of WeChat diffusion process. Cascade mass and breadth were more predictable than depth or Wiener Index. Prediction accuracy was stable at different observation sizes. By using case study with visualization, it was suggested that the content topic factor might intrinsically determine the pattern of structural and dynamical evolution. Our work enriched the knowledge of information diffusion mechanism on WeChat network, and facilitated the analysis and prediction practice. Future work includes obtaining more interesting diffusion patterns on comprehensive dataset, and improving prediction performance and by applying more effective features and more reliable learning algorithms.

Acknowledgements

We gratefully acknowledge the support provided social diffusion research group in WeChat Team, Tencent Co., Ltd. This paper was also partially funded by Shenzhen Municipal Development and Reform Commission, Shenzhen Environmental Science and New Energy Technology Engineering Laboratory, Grant Number: SDRC [2016]172.

References

[1] Maglio, P., C. Kieliszewski, and J. Spohrer. 2010. Handbook of Service Science. New York, NY: Springer.

[2] de Sola Pool I and Kochen M: “Contacts and influence”, Social networks, Vol.1, No.1, pp.5-51, 1978.

[3] Watts D. J. and Strogatz, S. H.: “Collective dynamics of small-world networks”, Nature, Vol.393, No.6684, pp.440-442, 1998.

[4] Barabasi A. L. and Albert R.: “Emergence of scaling in random networks”, Science, Vol.286, No.5439, pp.509-512, 1999.

[5] Fortunato S.: “Community detection in graphs”, Physics reports, Vol.486, No.3-5, pp.75-174, 2010.

[6] Pastor-Satorras R, Vespignani A.: “Epidemic spreading in scale-free networks”, Physical review letters, Vol.86, No.14, pp.3200, 2001.

[7] Wang C., Guan X., Qin T. and Zhou Y.: “Modelling on opinion leader’s influence in microblog message propagation and its application”, J. Softw, Vol.26, pp.1473–1485, 2015.

[8] Yang J., McAuley J. and Leskovec J.: “Community detection in networks with node attributes”, IEEE 13th International Conference on Data Mining, pp.1151-1156, Dallas, U.S.A., 2013.

[9] Hofman J. M., Sharma A. and Watts D. J.: “Prediction and explanation in social systems”, Science, Vol.355, No.6324, pp.486-488, 2017.

[10] Martin T., Hofman J. M., Sharma A. and Watt D.J.: “Exploring limits to prediction in complex social systems”, Proceedings of the 25th International Conference on World Wide Web, Montréal, Canada, pp.683-694, 2016.

[11] Vosoughi S., Roy D. and Aral S.: “The spread of true and false news online”, Science, Vol.359, No.6380, pp.1146-1151, 2018.

[12] Cheng J., Adamic L., Dow P. A., Kleinberg J. M., and Leskovec J.: “Can cascades be predicted?”, Proceedings of the 23rd international conference on World wide web, pp.925-936, Seoul, Republic of Korea, 2014.

[13] Guo R., Shaabani E., Bhatnagar A. and Shakarian P.: “Toward order-of-magnitude cascade prediction”, Proceedings of the 2015 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining, pp.1610-1613, Paris, France, 2015.

[14] Ugander J., Backstrom L., Marlow C. and Kleinberg J.: “Structural diversity in social contagion”, Proceedings of the National Academy of Sciences, Vol.109, No.16, pp.5962-5966, 2012.

[15] Lagnier C., Denoyer L., Gaussier E. and Gallinari P.: “Predicting information diffusion in social networks using content and user’s profiles”, European conference on information retrieval, pp.74-85, Berlin, Germany, 2013.

[16] Camerer and Colin F.: “Behavioral game theory: Experiments in strategic interaction”, Princeton University Press, 2011.