Quantifying Structural Patterns of Information Cascades

Chengxi Zang[1], Peng Cui[1], Chaoming Song[2], Christos Faloutsos[3] and Wenwu Zhu[1]
[1]Department of Computer Science, Tsinghua University, Beijing, China
[2] Department of Physics, University of Miami
[3] Computer Science Department, Carnegie Mellon University
zangcx13@mails.tsinghua.edu.cn, chaomingsong@gmail.com, christos@cs.cmu.edu,
{cuip,wwzhu}@tsinghua.edu.cn

ABSTRACT

Information cascades are ubiquitous in both physical society and online social media, taking on large variations in structures, dynamics and semantics. Although there has been much progress on understanding the dynamics and semantics of information cascades, little is known about their structural patterns. In this paper, we explore a large-scale dataset including 432 million information cascades with explicit records of spreading traces. We find that the structural complexity of information cascades is far beyond the previous conjectures. We first propose seven-dimensional metrics, which reflect size and spreading orientation aspects, to quantify the structural characteristics of millions of information cascades. Further, we analyze the correlations of these metrics, finding some brand new structure patterns of information cascades, potentially providing insights into intrinsic mechanisms governing information spreading in nature and new models to forecast as well as to impose good control over information cascades in real applications.

CCS Concepts

•**Networks** → **Online social networks;**

Keywords: *Social networks; Information cascades; Structures*

1. INTRODUCTION

Information cascades are ubiquitous phenomena in self-organized social systems, enabling the local individuals to have global senses, and thus playing important roles in technology dissemination as well as epidemic diffusion [1]. Due to the importance and complexity of this phenomenon, the information cascades have attracted considerable attention in recent years, ranging from the cascades of chain-letters [2] in physical society to the cascades of resharing in on-line social media platforms such as Facebook and Weibo [5]. Although the dynamics [6, 7] and semantics [3] of information cascades are well explored, a paucity of works examine the structural patterns of information cascades [4]. It remains an interesting problem to see how to quantify the structural patterns of information cascades, and how they look like in the metric space. Without an understanding of the structure patterns of cascades, modeling or forecasting the spreading process remains a challenge.

ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054214

One of the major reasons why this problem is seldom studied is the lack of data covering explicit and full traces of information cascades. In this paper, we collect 432 million information cascades in Tencent Weibo (t.qq.com), which is one of the largest microblog systems in China. This dataset includes *the full scale* information cascades generated during a 7-day period, and for each microblog we have the explicit records of its spreading traces.

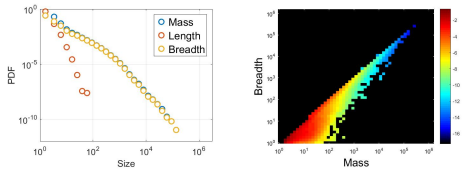
Through extensive observational study over the dataset, we find that real information cascades exhibit rich structures with large complexity. In order to quantify the complex structures, we propose seven-dimensional structural metrics to reflect the size and orientation aspects of information cascades. We then study the correlations of these metrics, and obtain insightful understanding on the structural patterns of cascades. For instance, we find that large cascades are either deep or wide (Fig. 1c), and the probability of generating large cascades with both large width and depth is quite small (Fig. 1d). We also find that besides the common branching-out cascades, there still exists converging-in cascades (Fig. 2a), where some users tend to retweet a microblog multiple times from his/her different followees, and the branching-out and converging-in structures tend not to co-exist in one cascade (Fig. 2c).

2. CASCADE STRUCTURE PATTERNS

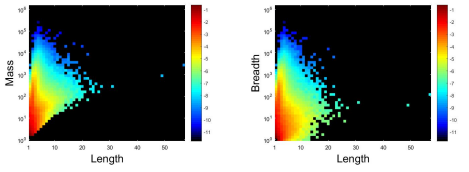
Cascade structure definition. The structure of a cascade $C = (V, E)$ is a directed graph in which each node $u \in V$ represents a user and each edge $(u, v) \in E$ represents that user v retweets user u's post. The user $u_o \in V$ who initializes the post is the original poster and all the other users are retweeters. An integer weight $w(u, v) \geq 1$ counts the number of edges from u to v, indicating the fact that v retweets u $w(u, v)$ times. A loop (u, u) is an edge that connects u to itself, indicating that user u retweets himself. Reciprocal edges r_{uv} are a pair of edges $(u, v) \in E$ and $(v, u) \in E$ where $u \neq v$, indicating the user u and v retweet each other.

Size. The size concept of a cascade is derived from the need of comparing a bigger to a smaller, a longer to a shorter, and a wider to a narrower. Thus, the size of a cascade (denoted as C) is measured by following three metrics: *i*) *Mass* N of a cascade refers to the amount of nodes in it, indicating that a cascade with more users is larger than the one with fewer users. *ii*) *Length* L of a cascade is the largest number of edges from the u_o to any other nodes through the spreading paths, indicating that a cascade with larger length value is longer than the one with smaller length value. *iii*) *Breadth* B of a cascade is the largest amount of nodes in it at the same depth, indicating that a cascade with larger width value is wider than the one with smaller width value.

Size distribution. Figure 1a plots the distributions of the three size metrics for the observed cascade. We observe fat-tailed nature of all the size metrics, implying that there exist very large cascades with respect to each size metric. For instance, in our dataset, the

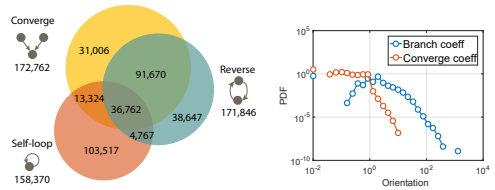


(a) Size (b) Mass-Breadth

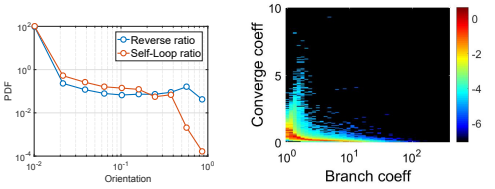


(c) Length-Mass (d) Length-Breadth

Figure 1: The size metrics for the empirical cascades.



(a) Venn diagram of converge, reverse and self-loop (b) Branch & Converge



(c) Reverse & Self-Loop (d) Branch-Converge

Figure 2: The orientation metrics for the empirical cascades.

biggest cascade which is also the widest cascade has mass value $1,414,815$ and breadth value $1,408,024$, and the longest cascade owes the length value 57. However, the average mass, breadth and length values are 5, 4 and 1 respectively.

Size correlation. Figures 1b-d plot the joint density profiles for the size metrics. We find strong positive correlation between mass and breadth as shown in Fig. 1b. Indeed, the correlation coefficient between the logarithmically transformed mass and breadth values is a strikingly high 0.99, indicating that the breadth accounts for a large proportion of mass. Figure 1c plots the joint distribution of length and mass, and we observe the biggest cascades are constrained to a relative small length, while the longest cascades are with moderate mass values. Figure 1d plots the joint distribution of length and breadth. We observe the majority of the cascades are wide and shallow, and there exist narrow and deep cascades. In contrast, it is difficult to find the very wide and deep cascades, or the very narrow and deep cascades.

Orientation. The orientation of a cascade measures to what extent that edges are directionally intertwined within the it. The orientation is characterized by following four metrics: *i*) *Branch coefficient* measures to what extent the edges in C spreading out to different nodes, characterized by the coefficient of variation (the ratio of the standard deviation to the mean) of out-degree distribution $p(k_{out})$ of C, where $k_{out}(u) = \sum_v \mathbb{1}\{(u,v) \in E\}$ is the out-degree of node u and $\mathbb{1}$ is the indicator function. A large branching coefficient value of C means the edges in C spread out from a couple of source nodes to a large amount of destination nodes, implying the orientation of spreading edges are fully random rather than spreading along a preferred direction. *ii*) *Converge coefficient* measures to what extent the edges in cascade C converging into one node, characterized by the coefficient of variance of in-degree distribution $p(k_{in})$ of C, where $k_{in}(v) = \sum_u \mathbb{1}\{(u,v) \in E \ \& \ u \neq v\}$ is the in-degree of node v. A cascade with large converging coefficient value indicates a large proportion of edges pointing to a couple of nodes, implying the information flows tend to converge into few users. *iii*) *Reverse ratio* measures to what extent the edges in cascade C pointing to the reverse direction, characterized by the ratio of the number of reciprocal edges to the total number of edges, i.e., $\frac{|\{(u,v)|(u,v)\in E \ \& \ (v,u)\in E \ \& \ u\neq v\}|}{|E|}$. *iv*) *Self-loop ratio* measures to what extent the edges in C starting and pointing to the same direction, characterized by the ratio of the number of nodes which have self-loop edge to the total number of nodes, i.e., $\frac{|\{u|(u,u)\in E\}|}{|V|}$.

Non-branching orientations. Existing studies of cascade focus mainly on the branching-out orientation, but we find other three orientations are also ubiquitous. Specifically, cascades containing at least one of the converge, reverse or self-loop orientations, shown in Fig. 2a, account for 20.0% of the total population. In addition, cascades usually show a combination of different spreading orientations. The Venn diagram in Fig. 2a shows the number of cascades with different orientation types and their logical relationships. In total, 9.2% of the total cascades have more than two orientation types, as shown in the overlapping region of Fig. 2. The branch coefficient distribution shows a bimodal distribution where two modes are near 0 and 2 respectively, implying that information flow in cascades tend to spread along one direction, or a moderate number of directions. In addition, very large values of branch coefficient do exist. In contrast, the distributions of converge coefficient, reverse ratio and self-loop ratio all peak near 0, a uniform-like distribution at a moderate value range, and followed by a fat-tail range at the large values, implying the prevalence of non-branching orientations and the existence of extreme cases (e.g. each node has a self-loop edge, or each edge has its reverse counterpart).

Orientation correlation. Further, we examine to what extent these spreading orientations can coexist. The branch orientation and converge orientation show non-coexistence relationship like two polarities. Figure 2d plots the heat map of branch coeff. vs. converge coeff. for each cascade. We find that large converge coeff. values only exist with small branch coeff. values, and vice versa.

Acknowledgments Supported by National Program on Key Basic Research Project, No. 2015CB352300; National Natural Science Foundation of China, No. 61370022, No. 61531006, No. 61472444 and No. 61210008, the fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, National Science Foundation under Grant No. CNS-1314632 IIS-1408924

3. REFERENCES

[1] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *15th ACM SIGKDD*, pages 199–208. ACM, 2009.

[2] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[3] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta. Large-scale high-precision topic modeling on twitter. In *the 20th ACM SIGKDD*, pages 1907–1916. ACM, 2014.

[4] G. Sharad, A. Ashton, H. Jake and W. Duncan The structural virality of online diffusion. In *Management Science*, INFORMS, 2015.

[5] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. In *2015 IEEE ICDM*,

[6] C. Zang, P. Cui, and C. Faloutsos. Beyond sigmoids: The nettide model for social network growth, and its applications. In *22nd ACM SIGKDD*, pages 2015–2024. ACM, 2016.

[7] T. Zhang, P. Cui, C. Song, W. Zhu, and S. Yang. A multiscale survival process for modeling human activity patterns. *PloS one*, 11(3):e0151473, 2016.