Uncovering Pattern Formation of Information Flow

Chengxi Zang
Department of Computer Science, Tsinghua University
zangcx13@mails.tsinghua.edu.cn
Department of Healthcare Policy and Research, Weill Cornell Medicine
chz4001@med.cornell.edu

Peng Cui
Department of Computer Science, Tsinghua University
cuip@tsinghua.edu.cn

Chaoming Song
Department of Physics, University of Miami
c.song@miami.edu

Wenwu Zhu
Department of Computer Science, Tsinghua University
wwzhu@tsinghua.edu.cn

Fei Wang
Department of Healthcare Policy and Research, Weill Cornell Medicine
few2001@med.cornell.edu

ABSTRACT

Pattern formation is a ubiquitous phenomenon that describes the generation of orderly outcomes by self-organization. In both physical society and online social media, patterns formed by social interactions are mainly driven by information flow. Despite an increasing number of studies aiming to understand the spreads of information flow, little is known about the geometry of these spreading patterns and how they were formed during the spreading. In this paper, by exploring 432 million information flow patterns extracted from a large-scale online social media dataset, we uncover a wide range of complex geometric patterns characterized by a three-dimensional metric space. In contrast, the existing understanding of spreading patterns are limited to fanning-out or narrow tree-like geometries. We discover three key ingredients that govern the formation of complex geometric patterns of information flow. As a result, we propose a stochastic process model incorporating these ingredients, demonstrating that it successfully reproduces the diverse geometries discovered from the empirical spreading patterns. Our discoveries provide a theoretical foundation for the microscopic mechanisms of information flow, potentially leading to wide implications for prediction, control and policy decisions in social media.

CCS CONCEPTS

• **Human-centered computing** → **Social networks**; **Social media**;

KEYWORDS

Social Media Analysis; Structure of Information Flow; Complex Pattern formation; Data-driven Branching Process

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3330971

ACM Reference Format:
Chengxi Zang, Peng Cui, Chaoming Song, Wenwu Zhu, and Fei Wang. 2019. Uncovering Pattern Formation of Information Flow. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330971

1 INTRODUCTION

Pattern formation is ubiquitous in nature [5, 34], ranging from physics [3, 33, 43] to biology [35, 46], from chemical reactions [55] to social interactions [1, 19, 30, 63, 64]. During the past years, much effort has been made towards modeling information flow across individuals in both physical society [10, 41] and online social media [12, 56, 64], aiming to enhance our understanding of the formation of complex social systems. Yet, little is known about the geometric patterns formed during the spreads of information [11, 20–22, 39, 49, 62]. Existing models such as epidemic models [23, 29, 48] and branching process [28, 36, 58] are prone to generating fanning-out or star-like patterns (Figure 1 C-D). However, real-world patterns of information flow seem to be much more complicated. For instance, Liben-Nowell and Kleinberg [41] found an unusual pattern with narrow and deep tree structure in the Internet chain-letter data. These findings raise a number of important questions:

- To what extent a complex spreading pattern can ever form?
- What are the underlying mechanisms governing the complex pattern formation of information flow?
- Can we generate realistic geometric patterns of information flow?

Answering these questions not only enhances our understanding of the formation of spreading patterns but also delivers computational tools to predict information flows, potentially leading to applications for dissemination of new technology [6, 7, 53, 62] and understanding the formation of public opinion [16, 18, 27, 59] or fake news [38, 57]. This paper tries to answer these questions.

A systematic study on the complexity of geometric patterns of information flow is missing partly due to the lack of reliable large-scale empirical datasets. Indeed, most studied datasets often lack explicit attributions tracking the information flow. For example, the Twitter dataset does not provide an explicit 'retweet' tag for each tweet, leaving difficulty in inferring the appropriate information flow [37, 54]. In this paper, we explore a large-scale social

media dataset which consists of 101 million users and 432 million information cascades among users within a 7-day period [65]. The presented data tracks all steps when a post is being created or forwarded, serving as one of the most accurate datasets to address aforementioned questions. Remarkably, we discover a broad class of geometries from the real-world spreading patterns (Examples shown in Figure 1 A), unveiling the complexity emerged during the pattern formation. Next, we introduce three key metrics to characterize geometric patterns, finding systematic discrepancies between existing models and empirical observations. Finally, we uncover three novel ingredients that govern the complex pattern formation of information flow, and then propose a stochastic process model based fundamentally on these three ingredients, which successfully captures the complex nature of geometric patterns observed in the real world.

It is worthwhile to highlight our contributions as follows:

- **Novel Findings:** Driven by a real-world information flow dataset with explicit spreading traces at large scale, we find complex geometric patterns of information flow and then propose metrics to quantify their geometric patterns in a three dimensional space.
- **Novel Mechanisms:** We find three key mechanisms which govern the formation of the complex geometric patterns of information flow, which are largely ignored by previous spreading models.
- **A Novel Model:** Based fundamentally on the mechanisms we find, we propose a stochastic process model which successfully captures the complex geometric patterns formed during the spreading of information flow.

2 RELATED WORK

Patterns of information flow in social media. How information spreads over social network is one of the core topics in both social science and computer science. Temporal/Dynamic patterns of information flow are studied empirically in literature [44, 51, 60, 62] etc. and in information cascading prediction literature [15, 40, 61, 67] etc. Influence modeling and maximization works [4, 14, 32, 52] try to make an information cascade reach audiences as many as possible, namely the final size of a cascade. However, these work largely ignores the structural patterns of information flow.

Seminal studies on the structural patterns of information flow includes [2, 21, 22, 39, 41, 49] etc. They find dominated star-like patterns of information flow [21] in social media like Twitter, and unusual pattern with narrow and deep tree structure in the Internet chain-letter data [41]. Besides, studies including [25, 50] try to infer the spreading structures from their temporal records. However, the lack of the explicitly spreading traces like the Twitter datasets and the limited number of data samples in the chain letter dataset [41] prevent us from a wholistic view of the structures of information flow. Indeed, with the access to new datasets of information flow in social media, recent empirical studies [42, 65] find rich complexities in the geometric patterns of information flow. However, principled metrics which quantify the complex geometric patterns of information flow are still missing. How to model the spreading process of information flow which captures these potentially complex geometric patterns is still largely unknown?

Spreading models. Current frameworks of modeling the spreading of information flow mainly fall into two categories: i) the epidemic model [31, 45, 48] which treats a spread of information flow as a contagion process between individual agents. Simple-contagion models [8, 29, 48] often assume independent pair-wise social interactions whereas complex contagion models [13, 17, 54] account for the possibility that a susceptible agent may get exposures from multiple infected neighbors simultaneously; ii) the branching process which assumes that each individual forwards the message to a set of offspring neighbors, where the total offspring size is drawn randomly from a predetermined offspring distribution function [24, 36, 58]. However, we find large discrepancy between these models and the empirical observations. These spreading models fail in capturing the complexities in the geometric patterns of information flow in the real-world.

3 QUANTIFYING THE GEOMETRIC PATTERNS OF INFORMATION FLOW

3.1 Dataset

We collected information flow data from Tencent microblog platform (t.qq.com), which records explicit information of each individual's activity when he/she creates or forwards a post. The underlying social network (followee-follower network) is reconstructed based on the observed forwarding activities (Sec. C in Supplementary Information). A cascade is constructed by following the information propagation starting from the original post and forwarded among followers. The dataset contains all the posts over a 7-day period between June 20, 2012 and June 26, 2012, consisting of 563, 331, 392 posts, 101, 802, 707 users, and 432, 101, 384 cascades.

3.2 Proposed Metrics

To quantify the geometric patterns of information flow, we use three representative metrics:

- *Mass* n that measures the total number of individuals involved in spreading,
- *Polarity* v that measures to what extent information flow is being directionally dependent, which implies different spreading tendency along different directions, characterized by the variance of the pair-distance among all the infected individuals in spreading (Detailed formula with examples in Supplementary Information Sec. A),
- *Outreach* d that measures the largest distance information flow can ever reach.

The proposed metrics (n, v, d) offer a three-dimensional space where different regimes of the space correspond to different geometries of the observed spreading patterns shown in Fig. 1 A. For instance, the observed large size patterns with narrow-and-deep structure correspond to large n, v, and d values. Indeed, a perfect chain pattern predicts the polarity $v \sim O(n^2)$ and the outreach $d \sim O(n)$, whereas a perfect star pattern predicts $v \sim O(0)$ and $d \sim O(1)$ independent of its mass n (see Sec. A in Supplementary Information). Figure 2 A-C enumerates all typical geometric patterns represented in the metric space.

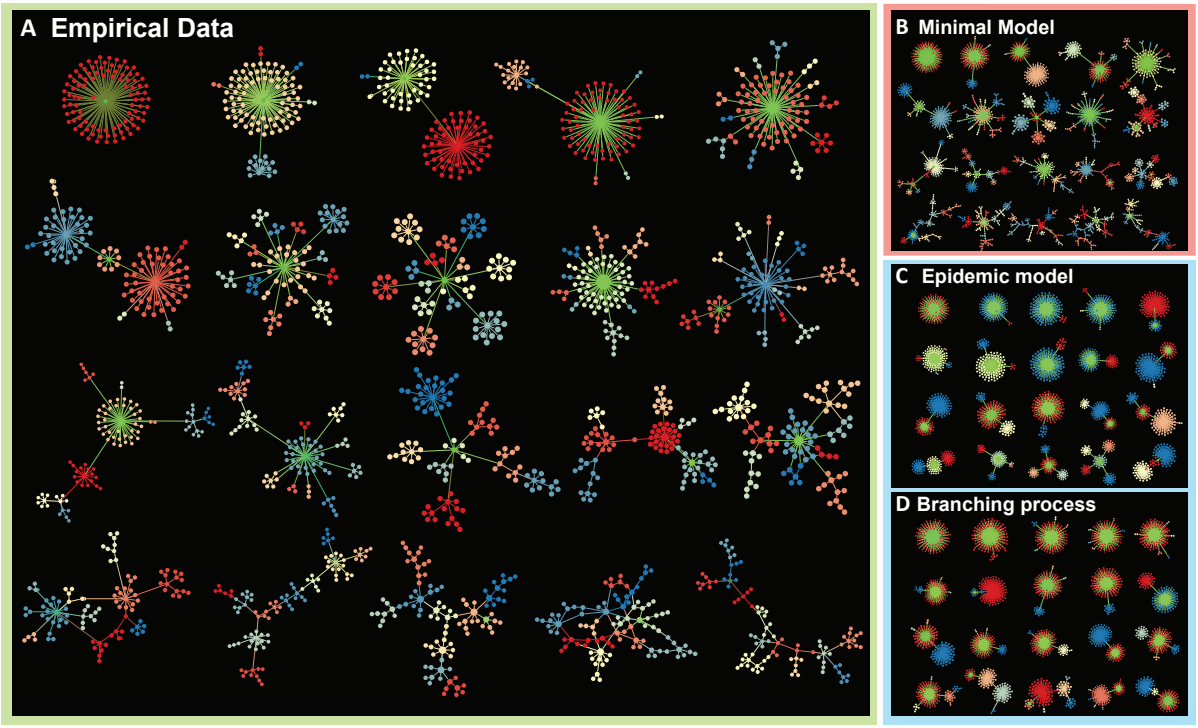

Figure 1: ***Illustrations of the geometric patterns of information flow for empirical data and different models.*** **(A) Twenty empirical patterns with size** 100 ± 3 **and increasing polarity value, each representing an information spreading starting at the center node colored in green. Nodes and links represent users and retweeting, respectively. Different colors correspond different community groups discovered by [9]. (B-D) Spreading patterns generated by (B) our proposed model, (C) the epidemic model and (D) the branching process model, respectively.**

3.3 Empirical Geometric Patterns

We measure n, v, and d for all 432 million spreading patterns (or cascades) observed in the dataset and plot the density profiles for n vs. v, n vs. d and d vs. v in Fig. 2 D-F respectively. We find that most spreading patterns fall into regions with small n, v, and d values, indicating that the vast majority of spreading patterns share a star-like pattern. We also observe the existence of large mass patterns (i.e., $n > 100$), implying the fat-tailed nature of the mass distribution of spreading patterns. Moreover, the polarity v and outreach d of these patterns vary over a large range of values, revealing the richness of geometric patterns in the real world. For instance, Figure 2 D-E show large v and d values for mass $n \sim 100$, indicating that the geometries of these patterns with moderate mass are most complex whereas ones with extremely large or small mass are relatively simpler. We also found in Fig. 2 F that the polarity and the outreach are positively correlated, yet there exists a large variance between these two metrics, implying rich complexities in empirical geometric patterns.

4 FAILURE OF EXISTING MODELS

Current frameworks of modeling the information flow mainly fall into two categories:

- the epidemic model [31, 45, 47, 48] which treats a spread of information flow as a contagion process between individual agents. Simple-contagion models [8, 29, 39, 48] often assume independent pair-wise social interactions whereas complex contagion models [13, 17, 54] account for the possibility that a susceptible agent may get exposures from multiple infected neighbors simultaneously;
- the branching process which assumes that each individual forwards the message to a set of offspring neighbors, where the total offspring size is drawn randomly from a predetermined offspring distribution function [24, 36, 58].

To demonstrate the discrepancy between existing models and the empirical observations, we perform simulations of the branching process model [36, 58] and two epidemic models [8, 26, 49], namely the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Recovered (SIR) model. We use the maximum likelihood estimation used in [24, 26] to evaluate the most likely modeling parameters of SIS from the real data, and generate 432 million cascades, each starting from the corresponding original poster observed in the empirical data [1]. Figure 1 C-D plots the spreading

[1] Here we plots the results of SIS model. The SIR model gives almost the same results as the SIS model

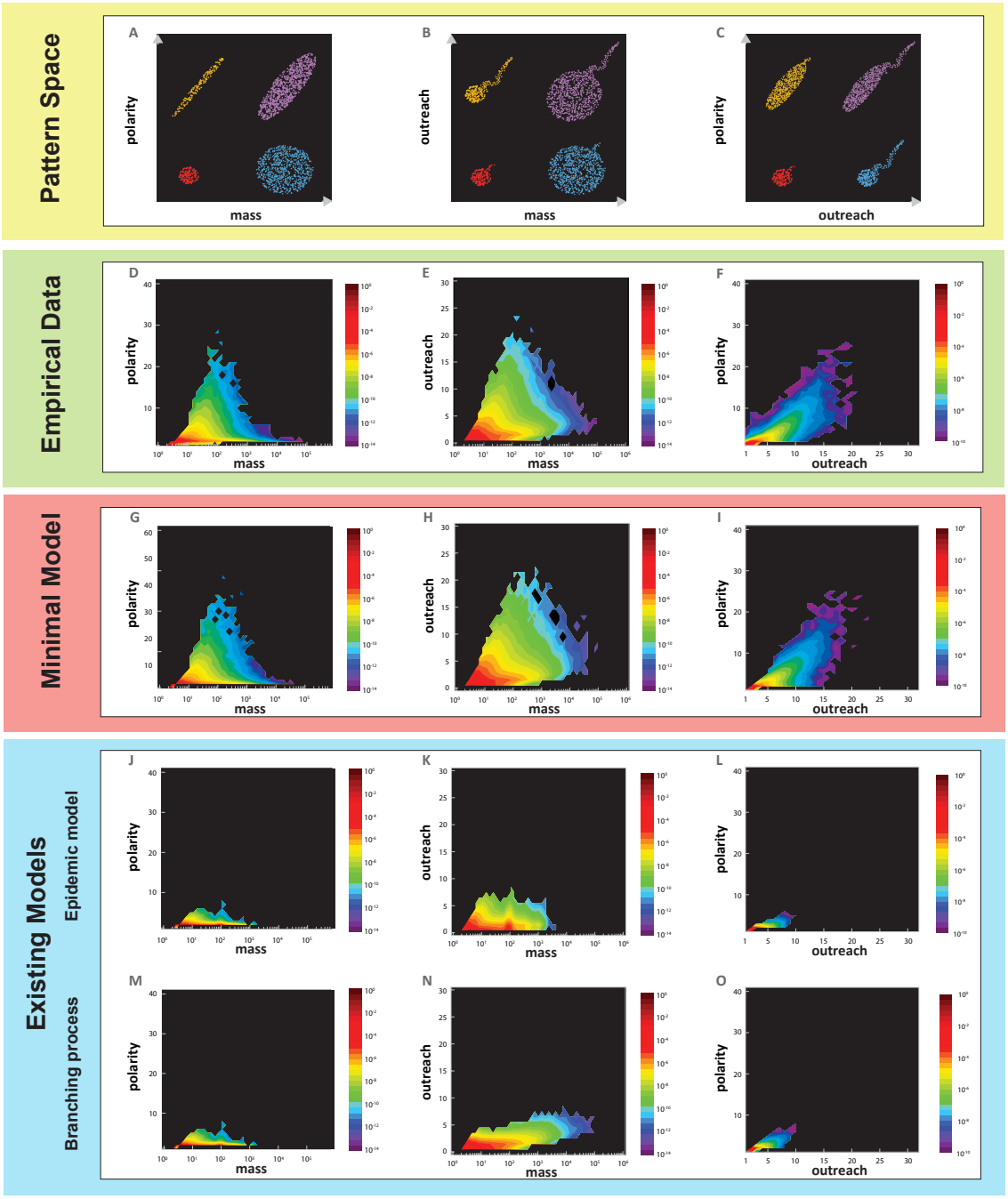

Figure 2: ***Quantifying the geometric patterns of information flow*** **via the three-dimensional metric space (mass, polarity, outreach). The projected two-dimensional space for (A) mass vs. polarity, (B) mass vs. outreach and (C) outreach vs. polarity, and we illustrate typical geometric patterns with typical metric values. Heat maps demonstrated two-dimensional probability density profiles (logarithmically transformed) for (D-F)** 432, 101, 384 **spreads extracted from the empirical data, (G-I) the proposed model, and the existing models including (J-L) the SIS model and (M-O) the branching process, respectively. All the models generate the same amount of spreads as the empirical dataset, and the modeling parameters are evaluated via maximum likelihood estimation from the empirical data.**

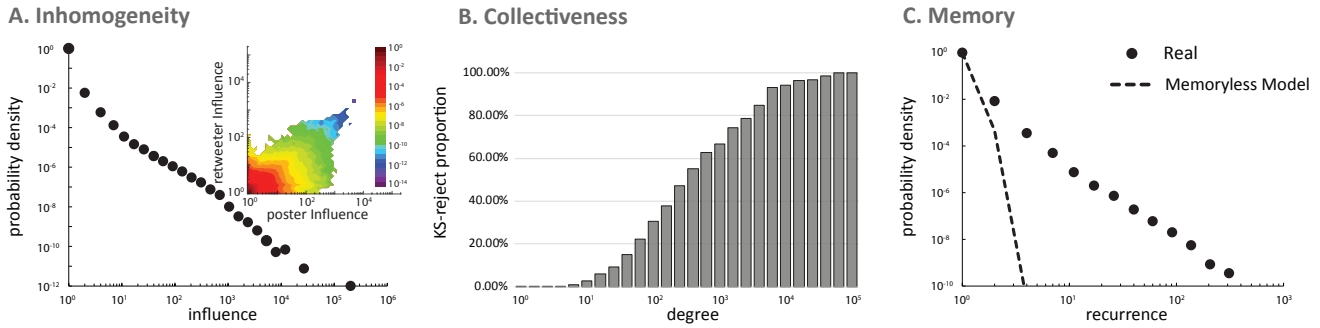

Figure 3: *Three ingredients governing complex pattern formation of information flow.* (A) The influence distribution on a log-log plot that demonstrates the heterogeneity across the userbase. The inset depicts the joint distribution of the influence for different roles, i.e., users acting as an original poster versus a retweeter, where the color represents the probability density. (B) Test of the collectiveness effect by evaluating the rejection rate of the null non-collective hypothesis. The bar plot shows the rejection proportion of users versus the users' degree, where we set 5% significance level for the two-sample Kolmogorov-Smirnov test. (C) The distribution of the recurrence, m, that measures the number of messages a user posts in a single spread, showing a fat-tailed distribution with a power law exponent $\gamma = 3.53 \pm 0.34$, and the dashed line plots the prediction of a null model without the memory effect and therefore is a narrow-tailed distribution.

patterns generated by the epidemic model and the branching process model respectively. Unlike the empirical data that reveals a rich set of spreading patterns, the existing models mostly generate star-like patterns. To quantify the difference, we measure mass, polarity, and outreach values for the generated spreading patterns by these models. Figure 2 J-O plots the joint distributions of the generated spreading patterns in our three-dimensional metric space. We find that, unlike the empirical data, for both models the polarity values are less than 10 and show little correlations with the mass (Figure 2 J&M), implying that these models can only generate simple patterns no matter how many individuals are involved in spreading. Similarly, we also observe that for both models the outreach values are very small, and the correlations between outreach and mass are very weak although there may exist some slight correlation for branching process (Figure 2 K&N). However, for both models polarity and outreach show strong positive correlations (Figure 2 L&O) in contrast with the empirical data (Figure 2 F).

In a word, existing models cannot capture the observed complex geometric patterns of information flow.

5 PROPOSED MODEL

5.1 Mechanisms

The failure of existing models implies the existence of unexplored mechanisms for the complex pattern formation of information flow which are not captured by these models. Indeed, we uncover that there are three essential ingredients governing complex pattern formation of information flow in the real world:

- *Heterogeneity.* In the real world, individuals have significantly different abilities to infect others. For instance, an individual with a larger number of connections often influences more neighbors than others. To quantify the difference between individuals' influence, we measure the number of neighbors infected by a user i (i.e., the offspring size b_i) for all the cascades in which i participated. The average offspring size $\langle b_i \rangle$ over all cascades characterizes the influence of individual i. Figure 3 A shows that the user influence $\langle b \rangle$ follows heavy-tailed distribution, demonstrating that individuals have highly inhomogeneous ability to infect others. Furthermore, individual plays different roles in information spreading. For example, a microblog user can act either as an original poster who initiates a cascade, or a retweeter who retweets the information from others. To quantify this, we measure each individual's influence for posting, $\langle b \rangle_p$, and for retweeting, $\langle b \rangle_r$, respectively. Figure 3 A inset plots the joint distribution of $\langle b \rangle_p$ versus $\langle b \rangle_r$, showing that despite the positive correlation between these two quantities there is a large variance between them. The above quantified evidence underpins the observation of heterogeneity in information spreading.
- *Collectiveness.* While the existing models, such as SIS model and SIR model, assume infection happens independently over each pair of individuals, we observe the evidence that infection dynamics often occur collectively, i.e., a group of users may be infected simultaneously, leading to a heavy-tailed offspring size distribution (Fig. 5 in Supplementary Information). In contrast, both SIS and SIR assume independent infection along each link and thus predict a binomial offspring size distribution. We apply two-sample Kolmogorov-Smirnov test to the observed offspring size distribution $p(b)$ of each user to measure the discrepancy between the predicted binomial distribution and the empirical data. Figure 3 B plots the rejection rate of the null hypothesis that the branching factor $p_i(b)$ of user i follows binomial distribution grouped by user's degree k (the number of friends), finding that the rejection rate increases with k, indicating the fact that larger degree nodes have stronger collective effect. For

instance, over 66% users with degree 1000 reject the null hypothesis, whereas the rejection rate increases to more than 94% for users with degree 10,000.

- *Memory*. Individual's spreading behaviors depend largely on his/her historical events, leading to a long range temporal correlation between information flows. To capture the memory effect at each individual level, we measure the recurrence, m, the number of times an individual appears in one cascade, and plot the recurrence distribution $p(m)$ over all cascades in Figure 3 C. We find that $p(m)$ follows a heavy-tailed distribution. In contrast, a memoryless model such as SIS model shows a narrow-tailed distribution. For example, we observe that a user appeared in a cascade more than 300 times, while the memoryless model only predicts the recurrence less than 3 times.

5.2 A Stochastic Process Model

Here we present a stochastic process model which incorporates all these observed ingredients to capture the complex pattern formation of information flow:

- Start from a single node i as an original poster, who randomly posts a seed of information, and then its offspring nodes can get infected because of posting.
- Draw the offspring size b randomly from the distribution p_i to determine the number of neighbors being infected afterwards, where p_i is the pre-determined modeling parameter that captures the offspring size distribution of the original poster i.
- Select randomly a set of b nodes from the neighbors of node i. The probability of a neighbor j to be selected is proportional to $w_{i,j,m}$, where m indicates the recurrence of the individual j. We assume the selection probability can be factorized as product $w_{i,j,m} = q_{i,j} \times \alpha_{j,m}$, where the modeling parameter $q_{i,j}$ captures the infective heterogeneity for the pair i and j, and the modeling parameter $\alpha_{j,m}$ captures the memory effect. A group of nodes are collectively selected to be infected in each step, which incorporates the collective effect.
- Repeat step 2 and 3 for the newly infected nodes until no node is further infected. Note that instead of using p_i, we use r_i as the offspring size distribution for the retweeters to incorporate the heterogeneity in spreading roles.

5.3 Parameter Inference

All the modeling parameters $(p_i, r_i, q_{i,j}, \alpha_{j,m})$ are estimated from the empirical data through maximum likelihood estimations.

Here we show the method on estimating the modeling parameters of our minimal model. We first estimate the offspring size distribution $p_i(b)$ and $r_i(b)$ for a user i when he/she acting as an original poster and a retweeter respectively. Let $f_i^p(b)$ ($f_i^r(b)$) be the number of cascades in which the user i acts as an original poster (a retweeter) with offspring size b, and $p_i(b)$ ($r_i(b)$) be the probability of user i with offspring size b when he/she acts as an original poster (a retweeter). Then the likelihood function of the original posters is

$$L_i^p(b) = \prod_b p_i(b)^{f_i^p(b)}, \tag{1}$$

and its corresponding log-likelihood function is

$$\ln L_i^p(b) = \sum_b f_i^p(b) \ln p_i(b). \tag{2}$$

To maximize the log-likelihood function subject to the constraint $\sum_b p_i(b) = 1$, we impose the Lagrange multiplier:

$$\Lambda(b, i, \lambda) = \sum_b f_i^p(b) \ln p_i(b) + \lambda(\sum_b p_i(b) - 1). \tag{3}$$

We require

$$\frac{\partial \Lambda}{\partial p_i(b)} = \frac{f_i^p(b)}{p_i(b)} + \lambda = 0, \tag{4}$$

and then leads to

$$\lambda = -\sum_b f_i^p(b), \tag{5a}$$

$$p_i(b) = -\frac{f_i^p(b)}{\lambda}. \tag{5b}$$

Finally, the maximum likelihood estimation (MLE) of $p_i(b)$ is the fraction of i with offspring size b when i acts as the original poster:

$$\hat{p_i(b)} = \frac{f_i^p(b)}{\sum_b f_i^p(b)}. \tag{6}$$

And $r_i(b)$ shares the same derivation procedure as $p_i(b)$. Thus, the MLE for $r_i(b)$ is:

$$\hat{r_i(b)} = \frac{f_i^r(b)}{\sum_b f_i^r(b)}. \tag{7}$$

The pairwise infection probability between user i and user j, denoted by $q_{i,j}$, is calculated by the ratio of $f_{i,j}$ to f_i, where $f_{i,j}$ is the total number of times user j retweeting user i's microblog, and f_i is the number of times i posting or retweeting microblogs:

$$q_{i,j} = \frac{f_{i,j}}{f_i}. \tag{8}$$

The memory factor $\alpha_{i,m}$ is proportional to the ratio of $c_{i,m}$ to c_i, where c_i is the number of cascades that user i participated, and $c_{i,m}$ is the number of cascades that user i participated for m times:

$$\alpha_{i,m} \propto \frac{c_{i,m}}{c_i}. \tag{9}$$

6 RESULTS

Figure 1 B illustrates spreading patterns generated by our proposed model, showing a good agreement with the empirical patterns as shown in Figure 1 A.

To further compare our proposed model with the empirical data quantitatively, we perform numerous simulations of information spreading over the empirical network (see Sec. C in Supplementary Information) where the spreading dynamic is determined by the proposed model.

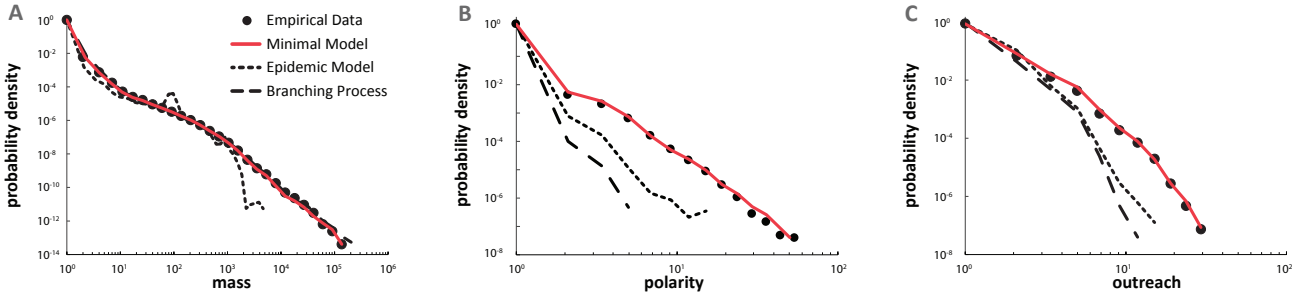

Figure 4: *Model evaluation.* Probability density functions for the three metrics (A) mass, (B) polarity and (C) outreach for spreads generated by our proposed model (red solid curve), the SIS model (black dotted curve) and the branching process (black dashed curve), compared with the empirical data (circles).

We measure the three-dimensional metrics, i.e., mass, polarity and outreach for all the simulated spreading patterns. Figure 4 A plots the mass distributions for the empirical data, the proposed minimal model, the epidemic model and the branching process model respectively, finding that both our model and the branching process fit well the empirical mass distribution while the epidemic model fails to capture the distribution's fat-tailed nature. Figure 4 B plots the polarity distributions for the empirical data and all the models. Our model predicts a heavy-tailed distribution [66] in line with the empirical observations. Previous models, however, only generate limited polarity values. Similarly, the outreach distributions shown in Figure 4 C also suggest a good agreement between our model and the empirical data while the existing models fail in reproducing the large outreach patterns.

Furthermore, our model captures not only the probability distribution of these metrics but also their correlations. Figure 2 G-I plots the density profiles for mass versus polarity, mass versus outreach, and outreach versus polarity for our model, respectively. Again, we find perfect agreements between the real patterns and the proposed model. In contrast, while the branching process predicts correctly the empirical mass distribution, it fails to capture the essential correlations between the mass and other metrics, i.e. polarity or outreach. Thus, our model captures the observed complex geometric patterns of information flow.

7 CONCLUSION

In summary, by exploring a large-scale dataset consisting of 432 million information cascades, we observe complex geometric patterns of information flow characterized by a three-dimensional metric space, finding systematic deviations from the prediction of the traditional epidemic models and the branching process. We find three ingredients, i.e., heterogeneity, collectiveness, and memory effect, which govern the complex pattern formation of information flow. Finally, we proposed a stochastic process model incorporating these ingredients that enables to reproduce the complex information flow patterns emerged in the real world. As our understanding on the mechanisms of information flow deepens with the emergence of increasingly detailed data, our discovery of the three fundamental ingredients and the proposed model suggest a possible basis for the future mechanistic understanding of the pattern formation of information flow. Our model can be potentially used to verify broad spreading mechanisms and phenomena, and can be potentially applied to prediction, control, and marketing scenarios in social media.

SUPPLEMENTARY INFORMATION

A POLARITY METRIC

Polarity v measures to what extent information flow being directionally dependent, implying different spreading tendency along different directions, which is characterized by the variance of the pair-distance among all infected individuals. Perfect star pattern and perfect chain pattern are two extremes with respect to polarity. A perfect star with mass n, denoted as S_n, consists of a central node and $n-1$ satellite nodes. The shortest path length (l) between any two nodes falls into two categories: $\binom{n-1}{2}$ satellite-to-satellite pairs with length 2 and $\binom{n-1}{1}$ satellite-to-center pair with length 1.

Then the polarity of a perfect star pattern is calculated by

$$v(S_n) = \frac{(2-\mu)^2\binom{n-1}{2}+(1-\mu)^2\binom{n-1}{1}}{\binom{n}{2}} = \frac{2n-4}{n^2} \tag{10}$$

where μ is the average distance between any two nodes in an undirected graph, calculated by

$$\mu(S_n) = \frac{2\binom{n-1}{2}+1\binom{n-1}{1}}{\binom{n}{2}} = 2 - \frac{2}{n}. \tag{11}$$

As for a perfect chain with mass n, denoted as C_n, there are $n-1$ pairs with length 1, $n-2$ pairs with length 2, ..., and 1 pair with length $n-1$. In order to get the closed form $v(C_n)$, we introduce:

$$A(n) = 1(n-1) + 2(n-2) + \ldots + (n-1)1 \tag{12}$$

$$B(n) = 1^2(n-1) + 2^2(n-2) + \ldots + (n-1)^2 1 \tag{13}$$

where $A(0) = B(0) = 0$ and $A(1) = B(1) = 1$. $A(n)$ and $B(n)$ can be transformed into recursive forms:

$$A(n) = A(n-1) + (n-1) + \ldots + 1 = A(n-1) + \frac{(n-1)n}{2} \tag{14}$$

$$B(n) = B(n-1) + (n-1)^2 + \ldots + 1^2 = B(n-1) + \frac{(n-1)n(2n-1)}{6} \tag{15}$$

By solving these two equations, we get:

$$A(n) = \frac{n(n^2-1)}{6} \tag{16}$$

$$B(n) = \frac{n^2(n^2-1)}{12} \tag{17}$$

Thus,

$$\mu(C_n) = \frac{A(n)}{\binom{n}{2}} = \frac{n+1}{3} \tag{18}$$

$$v(C_n) = \frac{B(n)}{\binom{n}{2}} - \mu(C_n)^2 = \frac{n^2-n-2}{18} \tag{19}$$

Indeed, a perfect chain pattern C_n predicts the polarity $v \sim O(n^2)$, whereas a perfect star pattern S_n predicts $v \sim O(0)$ independent with its mass n.

B OFFSPRING SIZE DISTRIBUTION

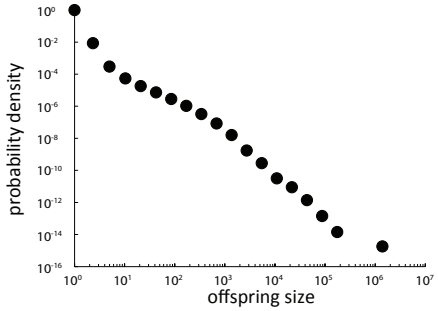

Figure 5: The offspring size distribution on a log-log plot.

We measure the number of infected neighbors for each individual user, i.e., the offspring size b, in each cascade. By aggregating 432 million empirical cascades we derive the offspring size distribution $p(b)$, as shown in Fig. 5. The fat-tailed nature of offspring size distribution implies the existence of a large group of users being infected collectively. For instance, although the average offspring size is 0.19, more than 10^6 users can be infected collectively.

C UNDERLYING SOCIAL NETWORK

We reconstruct the underlying social network from information flow records, i.e., we establish an edge between two individuals only if there exist spreading records between them. For instance, only if there is a piece of information spreading from user i to user j, j would be added to the possible offspring set of i. In this way, the directed relationships between any two individuals are captured. In total, we built offspring sets for all 101,802,707 users involved in our empirical dataset. To characterize the connectivity of the network, we measure the degree k, i.e., the number of offsprings, for each individual. Figure 6 plots the degree distribution of the underlying social network, showing that although the average degree is 0.6, very large hubs (e.g., $k > 10^7$) exist, implying the fat-tailed nature of the degree distribution.

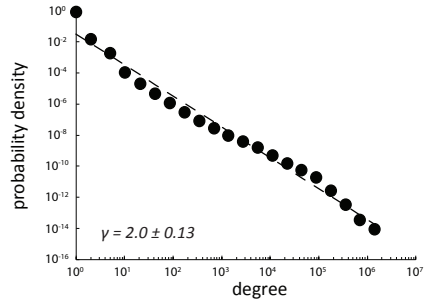

Figure 6: The degree distribution of the underlying social network. The dashed line in the log-log plot has slope -2.0.

Acknowledgments

The work is in part by supported by NSF IIS-1750326 and IIS-1716432, National Natural Science Foundation of China No. 61772304, No. 61521002, No. 61531006, No. U1611461, National Program on Key Basic Research Project (No. 2015CB352300), Beijing Academy of Artificial Intelligence (BAAI), the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, and the Young Elite Scientist Sponsorship Program by CAST.

REFERENCES

[1] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466, 7307 (2010), 761–764.
[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, and Mitul Tiwari. 2015. Global diffusion via cascading invitations: Structure, growth, and homophily. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 66–76.
[3] Christopher N Angstmann, Isaac C Donnelly, and Bruce I Henry. 2013. Pattern formation on networks with reactions: A continuous-time random-walk approach. *Physical Review E* 87, 3 (2013), 032804.
[4] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 65–74.
[5] Philip Ball. 2009. *Shapes: nature's patterns: a tapestry in three parts*. OUP Oxford.
[6] F Bass. 1969. A New Product Growth Model for Consumer Durables," Management Science, 15 (January), 215-227.(1980). *The Relationship Between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durables Technical Innovations,âĂİ Journal of Business* 53 (1969), 51–67.
[7] Frank M Bass. 2004. Comments on âĂIJA new product growth for model consumer durables the bass modelâĂİ. *Management science* 50, 12_supplement (2004), 1833–1840.
[8] Daniel F Bernardes, Matthieu Latapy, and Fabien Tarissan. 2012. Relevance of sir model for real-world spreading phenomena: Experiments on a large-scale p2p system. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 327–334.
[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
[10] Dirk Brockmann and Dirk Helbing. 2013. The hidden geometry of complex, network-driven contagion phenomena. *Science* 342, 6164 (2013), 1337–1342.
[11] Rui Castro, Mark Coates, Gang Liang, Robert Nowak, and Bin Yu. 2004. Network tomography: recent developments. *Statistical science* (2004), 499–517.
[12] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.
[13] Damon Centola and Michael Macy. 2007. Complex contagions and the weakness of long ties1. *Amer. J. Sociology* 113, 3 (2007), 702–734.
[14] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 199–208.
[15] Peng Cui, Shifei Jin, Linyun Yu, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Cascading outbreak prediction in networks: a data-driven approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge*

discovery and data mining. ACM, 901–909.
[16] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.
[17] Peter Sheridan Dodds and Duncan J Watts. 2005. A generalized model of social and biological contagion. *Journal of Theoretical Biology* 232, 4 (2005), 587–604.
[18] P Alex Dow, Lada A Adamic, and Adrien Friggeri. 2013. The Anatomy of Large Facebook Cascades.. In *ICWSM*.
[19] David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
[20] Mehrdad Farajtabar, Yichen Wang, Manuel Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. 2015. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*. 1945–1953.
[21] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan Watts. 2013. The structural virality of online diffusion. *Preprint* 22 (2013), 26.
[22] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. 2012. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*. ACM, 623–638.
[23] Alexander V Goltsev, Sergey N Dorogovtsev, JG Oliveira, and Jose FF Mendes. 2012. Localization and spreading of diseases in complex networks. *Physical review letters* 109, 12 (2012), 128702.
[24] Benjamin Golub and Matthew O Jackson. 2010. Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences* 107, 24 (2010), 10833–10836.
[25] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1019–1028.
[26] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 241–250.
[27] Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology* (1978), 1420–1443.
[28] Theodore E Harris. 2002. *The theory of branching processes*. Courier Dover Publications.
[29] Herbert W Hethcote. 2000. The mathematics of infectious diseases. *SIAM review* 42, 4 (2000), 599–653.
[30] Zhi-Qiang Jiang, Wen-Jie Xie, Ming-Xia Li, Boris Podobnik, Wei-Xing Zhou, and H Eugene Stanley. 2013. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences* 110, 5 (2013), 1600–1605.
[31] Matt J Keeling and Ken TD Eames. 2005. Networks and epidemic models. *Journal of the Royal Society Interface* 2, 4 (2005), 295–307.
[32] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.
[33] David A Kessler, Joel Koplik, and Herbert Levine. 1988. Pattern selection in fingered growth phenomena. *Advances in Physics* 37, 3 (1988), 255–339.
[34] AJ Koch and Hans Meinhardt. 1994. Biological pattern formation: from basic mechanisms to complex structures. *Reviews of Modern Physics* 66, 4 (1994), 1481.
[35] Shigeru Kondo and Takashi Miura. 2010. Reaction-diffusion model as a framework for understanding biological pattern formation. *science* 329, 5999 (2010), 1616–1620.
[36] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. 2010. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 553–562.
[37] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600.
[38] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
[39] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S Glance, and Matthew Hurst. 2007. Patterns of Cascading behavior in large blog graphs.. In *SDM*, Vol. 7. SIAM, 551–556.
[40] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. DeepCas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 577–586.
[41] David Liben-Nowell and Jon Kleinberg. 2008. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences* 105, 12 (2008), 4633–4638.
[42] Yunfei Lu, Linyun Yu, Tianyang Zhang, Chengxi Zang, Peng Cui, Chaoming Song, and Wenwu Zhu. 2018. Collective Human Behavior in Cascading System: Discovery, Modeling and Applications. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 297–306.
[43] Hernán A Makse, Shlomo Havlin, Plamen Ch Ivanov, Peter R King, Sona Prakash, and H Eugene Stanley. 1996. Pattern formation in sedimentary rocks: connectivity, permeability, and spatial correlations. *Physica A: Statistical Mechanics and its Applications* 233, 3 (1996), 587–605.
[44] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. 2012. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 6–14.
[45] Cristopher Moore and Mark EJ Newman. 2000. Epidemics and percolation in small-world networks. *Physical Review E* 61, 5 (2000), 5678.
[46] Akiko Nakamasu, Go Takahashi, Akio Kanbe, and Shigeru Kondo. 2009. Interactions between zebrafish pigment cells responsible for the generation of Turing patterns. *Proceedings of the National Academy of Sciences* 106, 21 (2009), 8429–8434.
[47] Mark EJ Newman. 2002. Spread of epidemic disease on networks. *Physical review E* 66, 1 (2002), 016128.
[48] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. *Reviews of modern physics* 87, 3 (2015), 925.
[49] Sen Pei, Lev Muchnik, Shaoting Tang, Zhiming Zheng, and Hernán A Makse. 2015. Exploring the complex pattern of information spreading in online blog communities. *PloS one* 10, 5 (2015), e0126894.
[50] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint:1105.0697* (2011).
[51] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (2014), 26–65.
[52] Manuel Gomez Rodriguez and Bernhard Schölkopf. 2012. Influence maximization in continuous time diffusion networks. *arXiv preprint arXiv:1205.1682* (2012).
[53] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.
[54] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 695–704.
[55] Alan Mathison Turing. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 237, 641 (1952), 37–72.
[56] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* (2012), 201116502.
[57] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[58] Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, and Albert-László Barabási. 2011. Information spreading in context. In *Proceedings of the 20th international conference on World wide web*. ACM, 735–744.
[59] Duncan J Watts. 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99, 9 (2002), 5766–5771.
[60] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 177–186.
[61] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. 2015. From Micro to Macro: Uncovering and Predicting Information Cascading Process with Behavioral Dynamics. In *IEEE International Conference on Data Mining, ICDM*.
[62] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 2015–2024.
[63] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2017. Long short memory process: Modeling growth dynamics of microscopic social connectivity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 565–574.
[64] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2018. On Power Law Growth of Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1727–1740.
[65] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu. 2017. Quantifying Structural Patterns of Information Cascades. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 867–868.
[66] Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2682–2691.
[67] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1513–1522.