Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph

DAVID R. BILD, YUE LIU, ROBERT P. DICK, and Z. MORLEY MAO, University of Michigan
DAN S. WALLACH, Rice University

Most previous analysis of Twitter user behavior is focused on individual information cascades and the social *followers* graph, in which the nodes for two users are connected if one follows the other. We instead study aggregate user behavior and the retweet graph with a focus on quantitative descriptions. We find that the lifetime tweet distribution is a type-II discrete Weibull stemming from a power law hazard function, that the tweet rate distribution, although asymptotically power law, exhibits a lognormal cutoff over finite sample intervals, and that the inter-tweet interval distribution is a power law with exponential cutoff. The retweet graph is small-world and scale-free, like the social graph, but is less disassortative and has much stronger clustering. These differences are consistent with it better capturing the real-world social relationships of and trust between users than the social graph. Beyond just understanding and modeling human communication patterns and social networks, applications for alternative, decentralized microblogging systems—both predicting real-word performance and detecting spam—are discussed.

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**]: Miscellaneous; J.4 [**Social and Behavior Sciences**]: Sociology

General Terms: Measurement, Human Factors

Additional Key Words and Phrases: Social network analysis, microblogging systems, decentralized network architectures

ACM Reference Format:
David R. Bild, Yue Liu, Robert P. Dick, Z. Morley Mao, and Dan S. Wallach, YYYY. Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Trans. Internet Technol.* V, N, Article A (YYYY), 24 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Quantitative modeling of Twitter usage is important both for understanding human communication patterns and optimizing the performance of other microblogging communication platforms. However, prior analysis is focused on the social graph [Kwak et al. 2010; Bliss et al. 2012; Teutle 2010; Gabielkov and Legout 2012; Ghosh et al. 2012] or on individual information cascades that represent a small fraction of all tweets [Suh et al. 2010; Java et al. 2007; Wu et al. 2011; Lotan et al. 2011; Galuba et al. 2010]. Descriptions of basic behaviors are missing from the literature. For example, the qualitative distributions of the number of *followers* and *friends* is available [Kwak et al. 2010], but not the distribution of tweet rates. Common factors of heavily retweeted tweets

This work was supported in part by the National Science Foundation under award TC-0964545 and in part by the Office of Naval Research under award 11536006.
Author's addresses: D. R. Bild (corresponding author), Y. Liu, R. P. Dick, and Z. M. Mao, Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109; email: drbild@umich.edu; D. S. Wallach, Department of Computer Science, Rice University, Houston, TX 77005.

DOI: http://dx.doi.org/10.1145/0000000.0000000

are known [Suh et al. 2010], but the propensity of users to retweet, i.e., distribution of retweet rates, is not. We begin to fill these gaps by considering user behavior *as a whole*, providing quantitative descriptions of the distributions of lifetime tweets, tweet rates, and inter-tweet times.

We are motivated by increasing interest in decentralized microblogging systems designed to protect user privacy and resist censorship. FETHER [Sandler and Wallach 2009], Cuckoo [Xu et al. 2010], and Litter [St Juste et al. 2011] reduce dependence on a single provider, while 1am [1am 2013] and Twister [Freitas 2013] are explicitly designed to avoid censorship and reprisal by government agencies. Designing a decentralized system capable of handling the message rates and volumes of Twitter is already a significant challenge and is nearly impossible without a good understanding of these usage patterns.

Given the complexity of these systems, understanding of the trade-offs in the performance and cost metrics—throughput, latency, energy consumption—is obtained through simulations, but such simulations are only as accurate as the data and models driving them. Consider fair allocation of network resources—fairness looks very different when the *expected* distribution of tweets is 80–20 power law and not uniform. Or, consider measuring delivery latencies, with messages queuing at intermediate nodes, a metric dependent on the (non-Poisson) arrival process, i.e., the inter-tweet duration distribution. Quantitative models of these basic behaviors are needed.

The underlying human behaviors should extend across communication platforms—tweet rates should mirror call rates in the telephone network and total lifetime tweets should mirror total lifetime contributions to Wikipedia or YouTube—suggesting that models of those behaviors [Wilkinson 2008; Seshadri et al. 2008; Candia et al. 2008] be used in proxy for microblogging design. However, our analysis of the Twitter data shows differing behavior, indicating possible faults in several of these models. Our results for Twitter should enable future work to identify or refine further commonalities in human communication.

Tweets generally travel via the explicit social *followers* graph [Kwak et al. 2010], which has been well-studied. Surprisingly, the *retweet* graph, in which a directed edge connects two users if the source has retweeted the destination, has received almost no attention. This *implicit* graph may be more relevant to information propagation in decentralized systems. A throughput-limited system needs some way of prioritizing messages. People are usually more selective in what they say than to whom they listen, so the retweet graph may better encode true interest and trust relationships among users. For example, 1am[1] does not support friend/follower relationships, so the retweet graph is the only available social graph. We conduct the first study of the retweet graph obtained from a 4-month sample of 10% of all tweets and compare it to the social followers graph.

These results have wide applicability. The quantification of communication behaviors and the social graph, beyond allowing direct comparison with other already-characterized platforms, enables the development of generative models explaining the underlying processes. In a more direct view, knowing the number of tweets, tweet rates, and inter-tweet times is sufficient for simulating and optimizing microblogging platform performance. The confirmation that the retweet graph is scale-free and small-word enables the generation of random retweet graphs for empirical evaluation. We focus on two such applications, the design of distributed microblogging systems and the detection of spammers using connectivity in the retweet graph.

[1] 1am [1am 2013] is a decentralized, geographic microblogging system in which messages are broadcast to users within radio range of the sender. Other users may re-post the message, extending its reach, but messages are not re-posted without endorsement by users.

We have the following main findings.

— The distribution of lifetime tweets is discrete Weibull (type-II), generalizing a power law form shown by Wilkinson for other online communities [Wilkinson 2008]. We conjecture that the Weibull shape parameter reflects the average amount of (positive or negative) feedback available to contributors. (Section 3)
— The distribution of tweet (and retweet) rates is asymptotically power law, but exhibits a lognormal cutoff over finite-duration samples. Thus, high tweet rates are much rarer in practice than the asymptotic distribution would suggest. We also discount a double Pareto lognormal (DPLN) explanation previously advanced in the context of call rates [Seshadri et al. 2008]. (Section 4)
— The distribution of inter-tweet durations is power law with exponential cutoff, mirroring that of telephone calls [Candia et al. 2008]. (Section 5)
— The retweet graph is small-world and scale-free, like the social followers graph, but less disassortative and more highly clustered. It is more similar than the followers graph to real-world social networks, consistent with better reflection of real-world relationships and trust. (Section 6)

In Section 7, we discuss the implications of these results for decentralized microblogging architectures and in Section 8 we consider using the structure of the retweet graph for spammer detection.

2. DATASETS

The Twitter API rate limits and terms of service prevent collection and sharing of a single complete tweet dataset suitable for all our queries [Watters 2011]. Our analysis focuses on a dataset containing 10% of all tweets sent between June and September 2012, but we supplement with sets from other researchers as necessary. This section summarizes these datasets and describes our main procedure for inferring population statistics from the 10% sample.

2.1. 2009 Social Graph

Kwak et al.'s 2009 crawl [Kwak et al. 2010] remains the largest and most complete public snapshot of the Twitter social followers graph, covering 41.7 million users and 1.47 billion relations. The data is dated, but still the best available. Repeating this crawl is infeasible under current rate limits, and feasible sampling strategies (e.g., snowball-sampling [Goodman 1961]) lead to results that are difficult to interpret [Lee et al. 2006]. We use this social graph snapshot for all comparisons with the retweet graph.

2.2. Lifetime Contribution Dataset

No tweet dataset is complete enough to compute lifetime contributions, the number of tweets sent before quitting Twitter, but the Twitter API exposes (subject to rate limits) the necessary information. We collected account age, date of last tweet, and total tweet count (as of June 2013) for 1 318 683 users[2] selected uniformly randomly from the 2009 social graph set.[3] 525 779 of these users were inactive, i.e., had not tweeted in the prior six months [Wilkinson 2008].[4] Their ages and tweet counts form the lifetime contribution set used in Section 3.

[2]We ran the rate-limited collection script for 19 h.
[3]The 2009 social graph dataset is the closest to a uniform random sample of Twitter users we could find. More recent sets are biased towards users that tweet more often.
[4]The creation dates of protected tweets are hidden, so all users with protected tweets were excluded.

Table I: 10% Sample (Gardenhose) Dataset

	10% Sample	**Actual Value**†
# of Tweets	4 097 787 713	41 256 584 408
# of Retweets	953 457 874	9 664 691 519
# of Tweeters	104 083 457	166 335 390
# of Retweeters	51 319 979	84 278 086
# of Retweetees	38 975 108	69 224 526

† Estimated using the described EM procedure.

2.3. SNAP Tweet Dataset

Computing inter-tweet intervals requires consecutive tweets—a random sample is insufficient.[5] For our inter-tweet distribution analysis in Section 5, we use a collection of 467 million tweets gathered by the SNAP team in 2009 [Yang and Leskovec 2011]. The full dataset is no longer publicly available per request from Twitter, but the authors kindly shared the inter-tweet metadata.

2.4. 10% Sample (Gardenhose) Dataset

Our primary dataset is a uniform random 10% sample[6] of all tweets (the "gardenhose" stream) sent in the four month period spanning June through September 2012. Table I shows the scope of the dataset, using the following definitions. A *tweeter* is a user who sends a *tweet*, an original message. A *retweeter* is a user who sends a *retweet*, forwarding a previous tweet. A *retweetee* is a user whose tweet was *retweeted*. Retweets were identified using both Twitter-provided metadata and analysis of the message contents for retweet syntax, e.g., "RT@".[7]

The sampled data poses a challenge for drawing quantitative conclusions about user behavior and the structure of the retweet graph. For many of the distributions we wish to quantify, the sample is biased towards users that tweeted more frequently. In fact, most users with fewer than ten tweets will not appear at all. Much prior work in the social network and graph analysis literature has focused on qualitatively characterizing the errors introduced by subsampling, motivated by quicker analysis [Lee et al. 2006; Son et al. 2012]. We instead give an approach to accurately estimate quantitative population statistics from the 10% random sample.

2.5. Estimating Population Distributions from the 10% Sample Dataset

For simplicity, we describe the method for a concrete problem: determining the distribution of tweets per user over our four month window. The method is trivially adapted to a variety of such problems, including multivariate joint distributions as in Section 6.3. This approach has been previously used in other fields [Duffield et al. 2005]. We wish to determine the number of users, f_i, with $i \in \mathbb{N}^+$ tweets, given the number of users, g_j, with $j \in \mathbb{N}^+$ tweets observed in the sample (remember that each tweet is observed with 10% probability). g_j includes some users from each $f_{i \geq j}$, with the binomial distribution $B_{0.1}(i, j)$ describing how the users in f_i are partitioned among the various $g_{j \leq i}$. Intu-

[5] A random sample would be sufficient if the process were Poisson, but it is not.

[6] More precisely, each tweet is included in this sample with 10% probability.

[7] Specifically, we used the following (Java) regular expression after lowercasing the tweet: `Pattern.compile("(?:^|[\\W])(?:rt|retweet(?:ing)?|via)\\s*:?\\s*@\\s*([a-zA-Z0-9_]1,20)(?:\$|\\W)")`

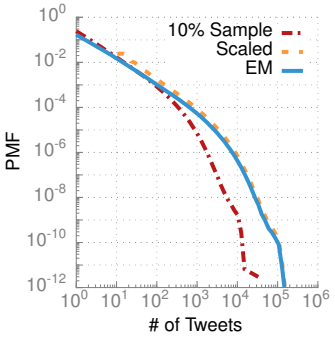

Fig. 1: Tweets per user for 10% sample, scaled correction, and EM correction.

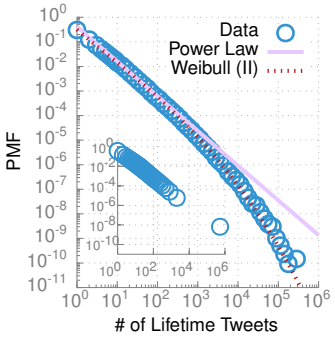

Fig. 2: Distribution of total lifetime tweets. See Table III for parameters.

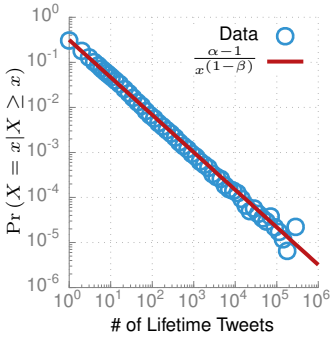

Fig. 3: Probability a user quits after x tweets or *participation momentum*.

itively, a good estimate $\hat{f}$ is that which maximizes the probability of the observation g, i.e., standard maximum likelihood estimation.

The corresponding likelihood function is not analytically tractable, so we employ an expectation maximization (EM) algorithm [Dempster et al. 1977; Borman 2009] to compute the estimate $\hat{f}$, summarized here. Let ϕ_i be the probability that a user sends i tweets conditional on at least one of them being observed and $c_{i,j}$ be the probability a user with i tweets has j of them observed conditional on $j \geq 1$ (i.e., the binomial probability conditioned on seeing at least one success). The log-likelihood function to maximize is $\mathcal{L}(\phi|f,g) = \sum_{1 \leq j \leq i} f_{i,j} \log\left(\phi_i c_{i,j}\right)$, where ϕ are the parameters to estimate and f and g are the hidden and observed variables, respectively. We compute the parameter estimate by iteratively selecting a new estimate ϕ^{k+1} that maximizes the *expected* likelihood under the previous estimate ϕ^k, i.e., $\phi^{k+1} \triangleq \arg\max_\phi \mathcal{Q}\left(\phi, \phi^k\right)$, where $\mathcal{Q}\left(\phi, \phi^k\right) \triangleq \mathsf{E}_{f|g,\phi^k}\left[\mathcal{L}(\phi|f,g)\right]$. This process is known to converge [McLachlan and Krishnan 2008]. Letting $\gamma \triangleq \sum_{1 \leq j} g_j$ be the total number of observed users, this maximization is solved using Lagrangian multipliers to yield $\phi_i^{k+1} = \frac{1}{\gamma}\,\mathsf{E}_{\phi^k}\left[f_i|g\right]$ or in matrix form for fast implementation on a computer, $\phi^{(k+1)} = \frac{1}{\gamma} \times \phi^{(k)} \times C \cdot \frac{g}{C^\top \cdot \phi^{(k)}}$. The hidden original frequencies are recovered from the final estimate $\hat{\phi}$ as $\hat{f}_i = \gamma \hat{\phi}_i \frac{1}{1 - B_{0.1}(i,0)}$.

Figure 1 shows the result using the distribution of tweets sent during our four-month collection window as an example. The correct distribution computed via the EM algorithm is substantially different, particularly in the lower decades, from the uncorrected or scaled (i.e., assuming that observing j tweets implies $10j$ were sent) distributions.

3. DISTRIBUTION OF LIFETIME TWEETS

Strong regularities in participation behavior have been observed across many online peer production systems, suggesting a common underlying dynamic. Wilkinson found that for Bugzilla, Essembly, Wikipedia, and Digg, the probability that a user makes no further contributions is inversely proportional to the number of contributions already made, suggesting a notion of *participation momentum* [Wilkinson 2008]. Huberman et al. observed the same in YouTube [Huberman et al. 2009]. We look for a similar effect in Twitter.

Table II: Contribution Momentum Exponents

Community	α	p	k_{min}
Essembly†	1.47	0.59	3
Digg†	1.53	0.64	15
Twitter	*1.54*	*0.96*	*12*
Bugzilla†	1.98	0.74	5
Essembly†	2.02	0.25	7
Wikipedia†	2.28	0.69	10
Digg†	2.40	0.04	15
Youtube‡	2.46	—	—

† [Wilkinson 2008]
‡ [Huberman et al. 2009]

Table III: Parameters for Lifetime Tweet Distributions Fit By MLE

Distribution			**Params**
Name	PMF		(MLE)
Power Law	$\frac{1}{\zeta(\alpha, x_{\min})} \cdot \frac{1}{x^{\alpha}}$	α	1.54
		$x_{\min}$	12.00
Type-II Discrete Weibull	$\frac{c}{x^{1-\beta}} \prod_{n=1}^{x-1} \left(1 - \frac{c}{n^{1-\beta}}\right)$	β	0.17
		c	0.32

We quantify contribution as the number of tweets sent,[8] so the lifetime contribution is the tweet count when the user becomes inactive. Following Wilkinson [Wilkinson 2008], a user who has not tweeted for six months (as of June 2013 when our lifetime contributions dataset was collected) is *inactive*.

Figure 2 plots the logarithmically-binned [Milojević 2010] empirical distribution. It is heavy-tailed, but decays more quickly in the upper tail than a true power law. The higher density in the last bin (∼200 000 tweets) is due to Twitter's rate limits of 1000 tweets per day and 100 tweets per hour,[9] because users that would occupy the upper tail (>200 000 tweets) are forced into this bin.[10] YouTube exhibits the same non-power law, upper tail cutoff [Huberman et al. 2009], consistent with a common dynamic underlying both systems.

3.1. Critique of Previously-Reported Power Law Behavior

Surprisingly, the cutoff does not match the strong power law evidence reported for Bugzilla, Essembly, Wikipedia, and Digg [Wilkinson 2008]. We believe those systems do contain a similar cutoff, but it was obscured by the analysis methods used. We observe three weaknesses in the prior analysis. First, the equal-count binning method[11] used obscures the upper tail behavior; logarithmic binning is preferable [Milojević 2010]. Second, maximum likelihood estimation, not binned regression, should be used for fitting [Clauset et al. 2009]. Finally, the goodness-of-fit should be computed against the empirical distribution function (Kolmogorov–Smirnov or Anderson–Darling test), not against binned data (the G-test) [Clauset et al. 2009].

The original datasets are unavailable,[12] so we tested our hypothesis by applying the same methods to our Twitter data. As expected, equal-count binning, shown in the inset of Figure 2, hides the known cutoff. The G-test for a power law fit by regression to the improperly binned data indicates a good match (Table II), despite the obvious mismatch in the real data. Clearly, these methods can obscure any underlying cutoff. Our results

[8] One could instead consider retweets, replies, or direct messages, but obtaining data for these is more difficult.
[9] We manually verified that the most-prolific accounts were tweeting at or near the rate limit.
[10] The rate limit means that the lifetime contribution distribution can be viewed as a censored [Johnson et al. 2005] version of the "natural" distribution, i.e., the natural rate for prolific users is only partially known.
[11] In equal-count binning, each bin is sized to contain the same number of samples and thus the same area under the density function. For B bins, the height of a bin b_i is computed as $B/w(b_i)$, where $w(b_i)$ is the width of b_i.
[12] Emails to the author bounced as undeliverable.

are consistent with Bugzilla, Essembly, Wikipedia, and Digg contributions containing the same cutoffs as Twitter and YouTube, but the original data would be needed to prove this hypothesis.

3.2. Lifetime Tweets Follow a Weibull Distribution

If the distribution is not power law, what is it? Examining the *hazard function*, or probability that a user who has made x contributions quits without another, provides the answer. Shown in Figure 3, the hazard function is an obvious power law. Wilkinson referred to this behavior in other online communities as *participation momentum* [Wilkinson 2008]; we will return to that interpretation later.

The power law hazard function $\frac{\alpha-1}{x^{1-\beta}}$ is that of the Weibull distribution,[13] for continuous support. For discrete support, the distribution with a power law hazard function is called a Type II Discrete Weibull[14] [Stein and Dattero 1984] and has mass function $\Pr(X=x) = \frac{\alpha-1}{x^{1-\beta}} \prod_{n=1}^{x-1} \left(1 - \frac{\alpha-1}{n^{1-\beta}}\right)$. A maximum likelihood fit to the lifetime contribution data yields $\beta = 0.17$ and $\alpha = 1.32$, as shown in Figure 2. The upper tail deviates slightly, which we attribute to Twitter's rate limit policy. Some users that would have tweeted more than ~200 000 times were artificially limited to fewer tweets, increasing the weight in that portion of the upper tail.

3.3. Interpreting the Hazard Function as Participation Momentum

Wilkinson [Wilkinson 2008] used a notion of participation momentum to explain the power law hazard function. For his assumed power law distribution, $C\frac{1}{x^{\alpha}}$, the hazard function is $\frac{\alpha-1}{x}$ and α can be seen as a metric for the effort needed to contribute. Higher required effort leads to a higher probability of quitting. Table II shows the α's for several systems. Intuitively, tweeting seems more taxing than voting on Digg stories but less so than commenting on Bugzilla reports. And indeed, we find that $\alpha_{\text{Digg}} < \alpha_{\text{Twitter}} < \alpha_{\text{Bugzilla}}$.

Alternatively, the hazard function might be more directly related to account age than total contributions. To reject this possibility, we compared the Kendall tau rank correlations [Kendall 1938] (a non-parametric measure of possibly *non-linear* correlation) between lifetime contributions, age, and average tweet rate (lifetime contributions/age). Unsurprisingly, age (i.e., longer life) correlates with increased lifetime contributions ($\tau = 0.4708$, $p = 0.00$, 95% CI $[0.4690, 0.4726]$). In contrast, the tweet rate is essentially uncorrelated with lifetime contributions ($\tau = -0.0067$, $p = 0.00$, 95% CI $[-0.0085, -0.0049]$), indicating that the momentum function is not driven by age. If it were, the correlation would be strongly positive because faster tweeters would generate more tweets in their (independently determined) lifetimes. The strong negative relationship between tweet rate and age ($\tau = -0.5687$, $p = 0.00$, 95% CI $[-0.5705, -0.5669]$) further supports this conclusion. The hazard rate is determined by the current total contributions, so users with higher tweet rates must have shorter lifetime ages.

The hazard function we observe ($\frac{\alpha-1}{x^{1-\beta}}$ instead of Wilkinson's $\frac{\alpha-1}{x}$) invites additional thought. The new parameter β ($\beta = 0$ in Wilkinson's model) models momentum gain—a higher β translates to more momentum gain per contribution. For example, one could imagine that β reflects a feedback effect. Positive (negative) viewer-generated feedback like retweets and replies in Twitter or comments and view counts in YouTube might accelerate (decelerate) momentum gains relative to systems without such visible

[13] The Weibull distribution is sometimes called the *stretched exponential*.

[14] The much more common Type I Discrete Weibull [Nakagawa and Osaki 1975] instead preserves the exponential form of the complementary cumulative density function.

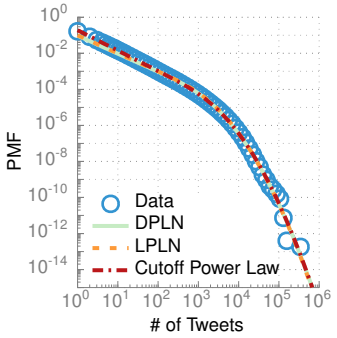

Fig. 4: Distribution of tweets per user for the four month period from June through September 2012. The DPLN, LPLN, and cutoff power law distributions differ in the lower tail.

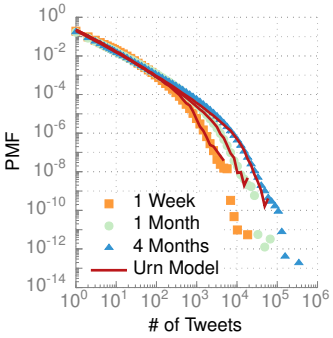

Fig. 5: Distribution of tweet counts showing the time-dependent cutoff. The asymptotic distribution is Pareto. Urn model traces were obtained by simulation.

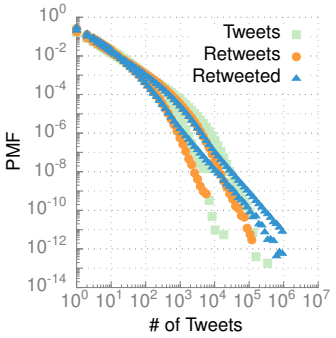

Fig. 6: Distributions of tweets sent (tweets), retweets sent (retweets), and times retweeted (retweeted) for the 1 week and 4 month sample periods.

feedback, like Digg votes or Wikipedia edits.[15] Refinement of this interpretation is a promising area for future work.

In summary, lifetime contributions in Twitter are driven by a power law hazard function $\left(\frac{\alpha-1}{x^{1-\beta}}\right)$ viewed as participation momentum. α reflects the effort needed to contribute and β the amount of feedback provided by system. The power law momentum leads to a Type II Discrete Weibull distribution for lifetime contributions. This dynamic holds across a variety of online communities [Wilkinson 2008; Huberman et al. 2009].

4. DISTRIBUTION OF TWEET RATES

The distribution of tweet rates is arguably the most important statistic for microblogging system design. An architecture designed for uniform messaging rates across the network will struggle with a heavy-tailed rate distribution. In this section, we describe an analytical model and generative mechanism for the rate distribution and reject a model previously proposed for telephone call rates. Although we are most interested in the tweet rate distribution, we model the easier-to-consider tweet count distribution. The former is easily recovered by factoring out the 4-month sampling duration.

4.1. An Analytical Approximation of the Tweet Rate Distribution

Figure 4 plots the logarithmically-binned empirical tweet distribution. It is heavy-tailed, consistent with other forms of authorship [Lotka 1926]. The tails form two different regimes meeting at $X = \sim 2000$, each heavy-tailed but with different exponents. We show in Section 4.3 that this phase change is a dynamic effect related to the sampling period (i.e., four months)—the crossing point increases with the square of the period.

A closed-form of the distribution makes simulating microblogging performance and comparing rates across communication systems much simpler. The forthcoming generative model in Section 4.3 is not analytically tractable, so we first describe an analytical

[15] Wilkinson's reported results are consistent with this hypothesis. The contribution types with the most visible feedback—Essembly and Digg submissions—show little support for a power law, with p-values of 0.25 and 0.04. $\beta > 0$ would explain the non-power law behavior. The distribution for YouTube by Huberman et al. also shows a cutoff [Huberman et al. 2009] consistent with a hazard function where $\beta > 0$.

approximation. Figure 4 suggests a cutoff power law, but the upper tail is heavier than the common exponential cutoff [Clauset et al. 2009]. Instead, the cutoff appears lognormal, suggesting the following density function,[16]

$$p(x) = cx^{-\beta}\Phi^c\left(\frac{\ln x - \mu}{\sigma}\right), \tag{1}$$

where Φ^c is the complementary CDF of the standard normal distribution and c is a normalizing constant. The maximum likelihood fit is shown in Figure 4, with $\beta = 1.13$, $\mu = 7.6$, $\sigma = 1.06$, and $c = 0.19$. The lognormal cutoff shape is seen by noting that $\Phi^c(z) \propto \mathrm{erfc}\left(\frac{z}{\sqrt{2}}\right)$ and $\mathrm{erfc}(z) \approx \frac{1}{\sqrt{\pi}}\frac{e^{-z^2}}{z}$ for $z \gg 1$, leading to the approximately lognormal form $\Phi^c\left(\frac{\ln x-\mu}{\sigma}\right) \propto \frac{\sigma}{\ln x-\mu}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$ for $\frac{\ln x-\mu}{\sigma} \gg 1$. The power law exponent in the lower tail is β, the phase change to the cutoff regime occurs at $\exp(\mu)$, and the steepness of the upper tail is controlled by σ.

4.2. The Distribution is Not Double Pareto–Lognormal

At first glance, Figure 4 appears to be Double Pareto-Lognormal (DPLN), a recently-discovered distribution that has found wide-spread popularity across many fields, perhaps due to its clear generative interpretation [Reed and Jorgensen 2004]. Seshadri et al. suggested its use for communication rates, specifically call rates in a cellular network, interpreting the generative process as evolving *social wealth* (a measure of how much social interaction a person has with others) [Seshadri et al. 2008]. However, in this section we show that the DPLN does not correctly capture the lower tail behavior of tweet rates (or call rates). In the next section, we describe a different mechanism to explain the shape.

We first summarize the origin of the DPLN distribution [Reed and Jorgensen 2004]. Given a stochastic process X evolving via Geometric Brownian motion (GBM) $\mathrm{d}X = \mu X + \sigma X\,\mathrm{d}W$ where W is the Wiener process and the initial state is log-normally distributed, $\log X_0 \sim \mathcal{N}(\nu, \tau^2)$, then X_t is also lognormally distributed, $\log X_t \sim \mathcal{N}(\nu + \frac{\mu-\sigma^2}{2}t, \tau^2 + \sigma^2 t)$. If the observation (or killing) time $t \triangleq T$ is exponentially distributed, $T \sim \mathrm{Exp}(\lambda)$, then the observed (or final) state has DPLN distribution, $X_T \sim \mathrm{DPLN}(\alpha, \beta, \nu, \tau)$, where $\alpha > 0$ and $-\beta < 0$ are the roots of the characteristic equation $\frac{\sigma^2}{2}z^2 + \left(\mu - \frac{\sigma^2}{2}\right)z - \lambda = 0$. Seshadri et al. [Seshadri et al. 2008] proposed that the number of calls made by an individual reflects an underlying *social wealth* that evolves via such a GBM. For an exponentially growing population, the age distribution of the sampled users is exponential and the resulting distribution of calls (or social wealth) will be DPLN. This model seems qualitatively reasonable for Twitter as well, but cannot capture the correct power law exponent in the lower tail (see Figure 4). The call distribution data exhibits a similar mismatch, undermining the model's suitability there as well.

The density function of $\mathrm{DPLN}(\alpha, \beta, \nu, \tau)$ is $f(x) = \frac{\beta}{\alpha+\beta}f_1(x) + \frac{\alpha}{\alpha+\beta}f_2(x)$, where $f_1(x) = \alpha x^{-\alpha-1}A(\alpha, \nu, \tau)\Phi\left(\frac{\ln x-\nu-\alpha\tau^2}{\tau}\right)$, $f_2(x) = \beta x^{\beta-1}A(-\beta, \nu, \tau)\Phi^c\left(\frac{\ln x-\nu+\beta\tau^2}{\tau}\right)$, $A(\theta, \nu, \tau) = \exp\left(\frac{\theta\nu+\theta^2\tau^2}{2}\right)$, and Φ and Φ^c are the CDF and complementary CDF of the standard normal distribution. f_1 and f_2 are the limiting densities as $\alpha \to \infty$ and $\beta \to \infty$, respectively, and are called the *right Pareto lognormal* and *left Pareto lognormal* distributions.

[16] We use a continuous model for simplicity. The integral data can be viewed as a rounded version of the product of the true tweet rate and sampling period.

Two observations stand out. First, the distribution is Pareto in both tails, with minimum slope of -1 in the lower. Second, the left Pareto lognormal form is nearly equivalent to our expression Equation 1, which differs only by accommodating lower tail exponents below -1. Figure 4 shows maximum likelihood fits of both the DPLN and left Pareto lognormal distributions. The lower tail is steeper than allowed by the DPLN ($-1.13 < -1$) and fits that model poorly. The call distribution data shows a similar mismatch. Although the DPLN is widely applicable, it does not model these communication patterns. Our model in the following section should better fit the call data [Seshadri et al. 2008] as well.

In the upper tail, both distributions fit equally well (i.e., a likelihood ratio test does not favor either fit). The data are insufficient to distinguish a lognormal from a power law upper tail, a common issue [Clauset et al. 2009]. We favor the lognormal form for Equation 1 because it is simpler (i.e., has fewer parameters) and most real world "power laws" exhibit some cutoff [Clauset et al. 2009].

4.3. An Urn Process Generating the Tweet Rate Distribution

In this section we develop an urn process to describe the tweet distribution in Figure 4. The phase change is a dynamic effect governed by the sampling period. As the period increases, the distribution approaches that of the lower tail—approximately Pareto with exponent -1.13. In practical terms, high-rate tweeters are much rarer in a finite sample than the asymptotic distribution would predict.

Figure 5 shows the distribution for three sample periods, illustrating the dynamic phase change. The lower tail extends further with the longer period. Degree distributions in growing networks evolve similarly. Although simple preferential attachment of new nodes leads to a straight power law [Barabási and Albert 1999], when existing nodes also generate new edges via preferential attachment, the distribution is double Pareto (with exponents -2 and -3) with a time-dependent crossing point ($k_c = [b^2 t(2+\alpha t)]^{1/2}$) [Barabási et al. 2002].[17] A similar model describes the tweet distribution.

Consider the evolution of the sample of tweets. Users join the sample upon their first tweet (during the sample period) and then continue to produce additional tweets at some rate. Discretize time relative to new users joining the sample, i.e., one user joins at each time step so there are t users at time t. Let $k(s,t)$ be the (expected) tweet count at time t for the user first observed at time s. Assume new tweets are generated at a constant average rate c, i.e., ct new tweets appear at each time step, distributed among existing users with frequency proportional to $A + k(s,t)^\alpha$. A is some initial attractiveness and α is the non-linearity of the preference [Dorogovtsev and Mendes 2002]. The resulting continuum equation[18] is $\frac{\partial k(s,t)}{\partial t} = (1+ct)\frac{A+k(s,t)^\alpha}{\int_0^t A+k(u,t)^\alpha \, \mathrm{d}u}$.

An analytical solution exists when $A = 0$ and $\alpha = 1$ [Dorogovtsev and Mendes 2001], but for the general case we resort to Monte Carlo simulations. Figure 5 shows the close match to the empirical density when $A = 1$ and $\alpha = 0.88$.[19] Assuming the power law form of the asymptotic density, $p(k) \propto k^{-\beta}$, the power law form of the rate distribution can be recovered. Taking λ as the tweet rate and noting that $\lambda \propto k^{-\alpha}$ when $k \gg A$, then

[17] In a network that allows self-edges, the exponents are -3/2 and -3 with crossing time $k_c \approx \sqrt{ct}(2+ct)^{3/2}$ [Dorogovtsev and Mendes 2000].

[18] We use the notation and continuous approximation of Dorogovtsev and Mendes [Dorogovtsev and Mendes 2001].

[19] Parameters were chosen by a coarse, manual exploration of the space. Fine-tuning might further improve the fit.

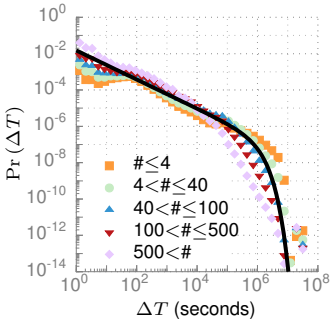

Fig. 7: The inter-tweet durations, grouped by number of tweets. Fit is a power law with exponential cutoff.

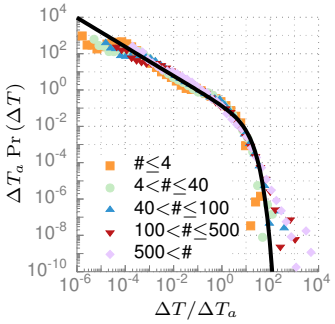

Fig. 8: The distributions collapse when scaled by the group's average duration, ΔT_a.

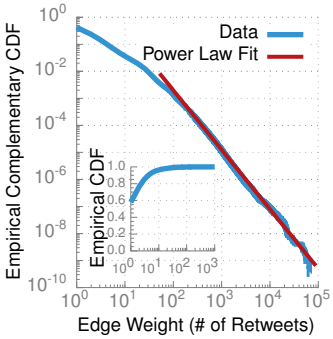

Fig. 9: Distribution of edge weights in the retweet graph, obtained via the EM method.

$p(\lambda) = p(k^{-\alpha}) \propto \frac{1}{\alpha}\lambda^{-\frac{-1+\alpha+\beta}{\alpha}}$. Thus, for α close to 1, the power law exponent recovered from Figure 4 slightly overestimates that of the tweet rate.

Going back to the analytical approximation of Equation 1, μ is related to ct by $\mu \approx 1.32\ln(ct) + 0.56$. $\beta = 1.13$ and $\sigma = 1.06$ are constants best determined by fitting.

4.4. Distributions of Retweeter and Retweetee Rates

The retweet and retweetee rates show a similar dynamic behavior in Figure 6. The retweet behavior differs only in the average rate c, which is about $2\times$ lower. The retweetee distribution exhibits two interesting differences. First, it extends further to the right, indicating that retweets of popular users outnumber tweets from prolific retweeters. Second, the slopes of the power law regimes are more consistent with a pure preferential attachment process (i.e., $\alpha = 1$). The retweetee rate comes directly from a preferential attachment process—initial retweets increase exposure, begetting additional retweets—and thus should match the linear form seen in other systems. The power law form of the tweet and retweet rates describes the underlying propensity to tweet, but without the same generative interpretation.

5. DISTRIBUTION OF INTERTWEET DURATIONS

Arrival processes in communication systems are traditionally assumed to be Poisson [Brown et al. 2005], but *per-individual* interval distributions for various activities including email, printing, and telephone calls are heavy-tailed [Barabási and Oliveira 2005; Harder and Paczuski 2006; Candia et al. 2008]. We show that this same behavior holds in Twitter, with our analysis mirroring that of Candia et al. for telephone calls [Candia et al. 2008] to enable easy comparison. The SNAP tweet dataset is used for this analysis.

We group the users by their total tweets to isolate the effects of differing tweet rates. Figure 7 plots the empirical distributions. Scaling by the group's average interevent time (Δt_a) collapses the distributions to a single curve, shown in Figure 8. This universal trait is also found in email and telephone systems [Goh and Barabási 2008; Candia et al. 2008]: the distribution is described by $\Pr(\Delta T) = \frac{1}{\Delta T_a} F\left(\frac{\Delta T}{\Delta T_a}\right)$, where $F(\cdot)$ is independent of the average rate. The best-fit cutoff power law is $\Pr(\Delta T) \propto (\Delta T)^{-\alpha} \exp\left(-\frac{\Delta T}{\tau_c}\right)$, with exponent $\alpha \approx 0.8$ and cutoff $\tau_c \approx 8.1\,\mathrm{d}$, shown as the black line in Figure 8. Here, ΔT_a is the average across the entire population.

6. CHARACTERISTICS OF THE RETWEET GRAPH

The natural and explicit network in Twitter—the social graph in which a directed edge represents the *follower* relationship—has been well-studied. Kwak et al. first reported on basic network properties like degree distribution, reciprocity, and average path length [Kwak et al. 2010], and later works have studied these and other characteristics in more detail [Bliss et al. 2012; Teutle 2010; Gabielkov and Legout 2012; Ghosh et al. 2012]. However, an alternative, implicit network—the retweet graph in which a directed edge indicates that the source retweeted the destination—has been largely neglected. We conduct the first characterization of the retweet graph and confirm that it, like many real-world networks, is small-world and scale-free. The reported metrics are useful for generating random retweet graphs using general parametric models like R-MAT [Chakrabarti et al. 2004] ($a = 0.52$, $b = 0.18$, $c = 0.17$, $d = 0.13$) or other more specific generative models [Bollobás et al. 2003].

We pay particular attention to contrasting the social following[20] and retweet graphs. Intuitively, they should be similar because retweets are usually sent by followers. However, we conjecture that the retweet graph more closely models the real-world social and trust relationships among users, because it derives from a more forceful action—not just listening to others' ideas, but actively forwarding them to one's own friends. Using the follower graph as a trust proxy has been proposed for applications ranging from spam filtering [Benevenuto et al. 2010; Song et al. 2011; Yang et al. 2012] to Sybil detection [Yu et al. 2008b; Yu et al. 2008a]. We conjecture that the retweet graph is a better choice and provide some supporting evidence. Full treatment of this conjecture is a promising avenue for further study.

6.1. Analyzing a Random Subsample of the Retweet Graph

The retweet graph is constructed from our largest dataset, the 10% sample, and contains 59 659 366 vertices and 426 584 484 edges. This sample does not include all edges. Rather, an edge is included with probability proportional to the number of retweets sent along it. However, 60% of edges have a single retweet and 98% have fewer than 10 (see Figure 9), so for simplicity we assume each edge is included with 10% probability. Many measured properties in an edge-sampled graph differ from the original graph. When possible, we use the EM-based method from Section 2.5 to correct our results. When not, we estimate the errors using the literature on sampled graphs [Lee et al. 2006; Son et al. 2012; Stumpf et al. 2005].

6.2. Degree Distributions

We begin with the in- and out-degree distributions. The in-degree k^i_{in} of a node i is the number of unique users who retweeted i and the out-degree k^i_{out} is the number of unique users retweeted by i. The average in-degree $\langle k_{\text{in}} \rangle \triangleq N^{-1} \sum_{i \in V} k^i_{\text{in}} = 88.4$ and the similarly-defined average out-degree $\langle k_{\text{out}} \rangle = 74.3$. V is the set of nodes, and N is their cardinality. In reality $\langle k_{\text{in}} \rangle = \langle k_{\text{out}} \rangle$; the observed difference is an artifact of the EM-based population estimation. The degree standard deviations are $\sigma_{\text{in}} = 4187.3$ and $\sigma_{\text{out}} = 228.4$. Higher in-degree variance is expected because, as with real-world networks [Son et al. 2012], *popularity* (the number of users who retweeted an individual) is more variable than *prolificity* (the number of users an individual retweeted).

The distributions, shown in Figure 10, are similar to the social following graph [Kwak et al. 2010]. Both are heavy-tailed and exhibit the same two-phase power law common to such networks. Similar to the tweet rate distribution (Section 4.3), the two phases are

[20]The social following graph is simply the social follower graph with the edge direction reversed to match that of the retweet graph.

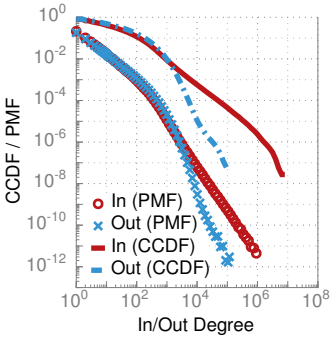

Fig. 10: Retweet graph degree distributions, showing double-Pareto behavior.

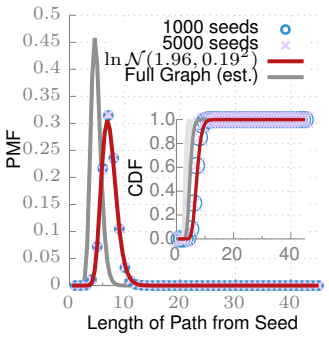

Fig. 11: Path length distribution in 10% edge-sampled retweet graph.

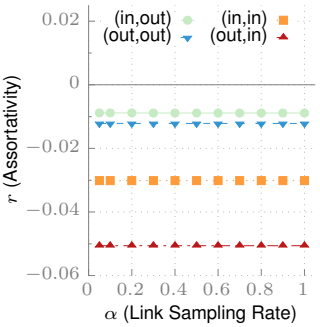

Fig. 12: Edge-sampling does not affect directed assortativities r.

a dynamic effect arising from two forms of evolution in the graph [Barabási et al. 2002; Dorogovtsev and Mendes 2001]—the addition of new nodes and preferential attachment of new edges between existing nodes. The outgoing (incoming) node i for a new edge is selected with relative probability $d_{\text{out}}(i) + \delta_{\text{out}}$ ($d_{\text{in}}(i) + \delta_{\text{in}}$), where δ_{out} and δ_{in} are the initial attractiveness constants and $d(\cdot)$ returns the node degree. Bollobás et al. elucidate this process for a general context [Bollobás et al. 2003].

The power law exponents are determined by δ_{out} (δ_{in}). The lower tails are similar with $\alpha \approx 1.3$. In the upper tail, $\alpha_{\text{out}} = 3.75$ and $\alpha_{\text{in}} = 2.2$. a_{in} matches the followers graph (2.3) [Kwak et al. 2010] and is in the range of most real-world networks (2–3). α_{out} exceeds that range because prolificity is not inherently preferential like popularity.

6.3. Reciprocity

Reciprocity is the fraction of links that are bidirectional. Many social networks have high reciprocity—most relationships are bidirectional (68% in Flickr [Cha et al. 2009] and 84% on Yahoo! 360 [Kumar et al. 2006]). In the Twitter follow graph, reciprocity is lower at just 22.1% [Kwak et al. 2010]. If retweeting is more discriminating than following, the retweet reciprocity should be lower. Indeed, it is just 11.1%.[21] Higher reciprocity in the follower graph may stem from the popularity of follow-back schemes in which a user, in an attempt to gain followers, promises to follow back anyone who follows him. The low reciprocity suggests that using the retweet graph as a proxy for trust is promising. Although a malicious node can establish many outgoing links, it has little control over the incoming links.

6.4. Average Shortest Path Length (Degree of Separation)

The real-world human social network has a small average shortest path length (APL) of about six, shown most famously by Stanley Milgram [Milgram 1967; Travers and Milgram 1969]. Many online networks are similar [Watts and Strogatz 1998; Leskovec and Horvitz 2008], but the social followers graph is denser with an APL of 4.12 [Kwak et al. 2010]. Kwak et al. attribute this difference to Twitter's additional role as an information source. Edges are more dense because users follow *both* social acquaintances and sources of interesting content.

[21] We estimated the distribution of all non-zero pairwise edge weight tuples (the number of retweets in both directions) from the 10% sample using the EM algorithm. The fraction that are non-zero in both directions is the reciprocity.

We determined the path length distribution of the 10% edge-sampled graph by computing all shortest paths for both 1000 and 5000 random starting nodes. The obtained distributions (shown in Figure 11) overlap, indicating a sufficient sample size. Lee et al. showed that edge sampling increases the APL by 1.5–$3\times$ (the gray range in the inset plot) depending on the graph structure [Lee et al. 2006]. We use $1.5\times$, determined by sampling the followers graph,[22] to estimate the full distribution (grey line in plot). The estimated APL is 4.8 and the 90th-percentile or *effective diameter* [Palmer et al. 2001] is 8.5. The difference from the followers graph is within estimation error.

The best-fit distribution (solid line in plot) is log-normal[23] with $\mu = 1.5$ and $\sigma = 0.27$. This differs from undirected Erdös-Rényi (ER) graphs, for which the limiting distribution is Weibull [Bauckhage et al. 2013], but we do not know of similar theoretical results for directed graphs.

6.5. Assortativity (Node Degree Correlation)

Degree assortativity—the tendency of nodes to connect with others of similar degree—summarizes the structural characteristics that in part determine both how content (e.g., retweets or disease) spreads and resilience to node removal [Newman 2002]. In an assortative network, content easily propagates through connected components of tightly clustered, high degree nodes that are resistant to node removal, but may not reach the low degree boundary of the network. Conversely, a disassortative network has larger connected components so content propagates further, but can be partitioned by the removal of a high degree node.

For undirected networks, assortativity is simply the Pearson correlation between the degrees of adjacent nodes.[24] The concept generalizes to directed networks by considering all possible directional degree pairs as separate assortativity metrics [Foster et al. 2010], $r(in, in)$, $r(in, out)$, $r(out, in)$, $r(out, out)$, again using the Pearson correlation $r(\alpha, \beta) \triangleq \frac{\langle k_\alpha^i k_\beta^j \rangle - \langle k_\alpha^i \rangle \langle k_\beta^j \rangle}{\sigma_{k_\alpha} \sigma_{k_\beta}}$, where $\alpha, \beta \in \{in, out\}$, k_α^i (k_β^j) is the α-degree (β-degree) of source (destination) node i (j), the averages $\langle \cdot \rangle$ are taken over all directional edges $(i \to j)$, and σ_{k_α} (σ_{k_β}) is the variance of the α-degree (β-degree).

We characterize and contrast these metrics for both the Twitter social following graph [Kwak et al. 2010] and retweet graph. Edge sampling impacts the degrees of all nodes identically and thus does not affect assortativity (see Figure 12) [Lee et al. 2006].

Figure 13 plots the assortativities for both networks. Although most real-world social networks are assortative [Newman 2002], online social networks are instead disassortative [Hu and Wong 2009]. The social followers graph is no exception, showing weak disassortativity across all measures. In contrast, the retweet graph is more assortative across all measures. It is near-neutral for both $r(in, \cdot)$ metrics, indicating independence between one's own retweet behavior and the number of retweets received. This is consistent with the graph containing useful trust information, because a user cannot influence the quantity of retweets received by selectively retweeting popular ($r(in, in)$) or prolific ($r(in, out)$) users. The high (out, out) assortativity is more consistent with real-world social networks and indicates that prolific retweeters retweet each

[22] The 2009 crawl [Kwak et al. 2010] is complete, so we compared the true statistic against that of a 10% subsample.

[23] We compared with the Weibull, Gumbel, Fréchet, and the encompassing generalized extreme value distributions.

[24] Litvak and van der Hofstad recently showed that the Pearson-based assortativity metric decreases with network size, complicating comparisons across networks of different size [Litvak and van der Hofstad 2013]. They recommend using a rank-order correlation, like Spearman's rho, instead. We are comparing with networks of similar size (i.e., the follower's graph), so Pearson's metric is acceptable and facilitates comparison with previously published metrics for similarly-sized graphs.

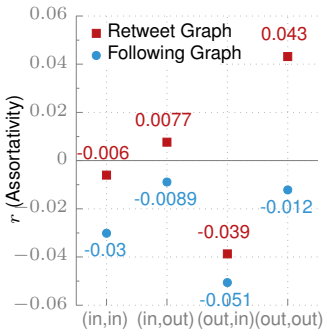

Fig. 13: Directed assortativity r of retweet and following graphs.

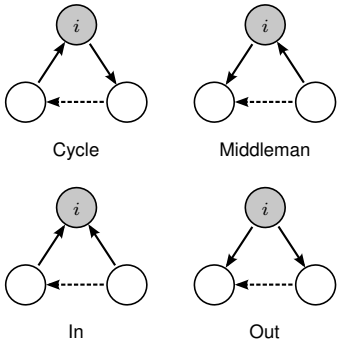

Fig. 14: Four types of directed triplets for clustering coefficient analysis.

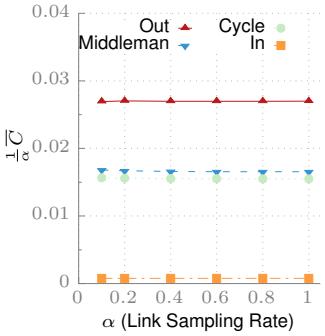

Fig. 15: The estimator $\widehat{C} \triangleq \frac{1}{\alpha}\overline{C}$ is accurate across sampling rates.

other. Interestingly, they are not tightly clustered (or the (in, out)-assortativity would be higher).

In Twitter, tweets propagate to followers, so the social graph disassortativity is helpful. The connected component is larger and tweets disseminate further more quickly. Increased susceptibility to node failure is acceptable in a centralized system. In a decentralized system that relies more heavily on the retweet graph structure for propagation, e.g., 1am [1am 2013], the resilience to node failure implied by its neutral and positive assortativities would instead be helpful.

6.6. Clustering Coefficient

A clustering coefficient quantifies the tendency of neighboring nodes to form highly connected clusters. Many real-world networks exhibit tighter clustering than would be expected in similar random graphs [Watts and Strogatz 1998]. We consider the *global clustering coefficient*,[25] defined for undirected graphs as $C \triangleq \frac{3N_\triangle}{N_3}$, where N_3 is the number of open *or* closed triplets (three vertices connected by two or three edges) and $N_\triangle$ is the number of closed triplets (3-vertex cliques). Unlike the alternative *local clustering coefficient*, this definition is suitable for networks with isolated nodes [Kaiser 2008]. In essence, C gives the probability that any two neighbors of a node are themselves connected.

Following the approach introduced by Fagiolo for the local clustering coefficient [Fagiolo 2007], we extend the metric to directed graphs by separately considering the four types of directed triplets, shown in Figure 14. The four clustering coefficients $C_{\beta\in\{\text{cycle, middleman, in, out}\}}$ are the fraction of β-triplets that are closed.

An estimator of the population clustering coefficient of an α-edge sampled graph ($\alpha = 0.1$ for us) is $\widehat{C} \triangleq \frac{1}{\alpha}\overline{C}$, seen by noting that a triplet is included in the sample with probability α^2 and as a closed triplet with probability α^3. This estimate is biased, because the triplets are not independent and edges could be concentrated towards open (or closed) triplets. In practice however, it performs well on large samples, as shown in Figure 15 for the social following graph.

Figure 16 plots the results for both the social and retweet graphs. The former has low clustering, but clustering in the retweet graph is significant for all metrics except *in*. *Cycle* is the only fully (transitively) connected triplet type, and thus *cycle*-clustering

[25] Sometimes called the *transitivity* or *transitivity ratio*.

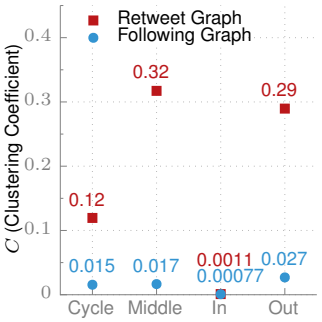

Fig. 16: Clustering coefficients for the following and retweet graphs.

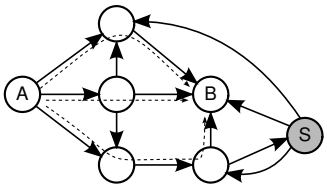

Fig. 17: Example retweet graph showing that non-spammer B is better connected to A than spammer S.

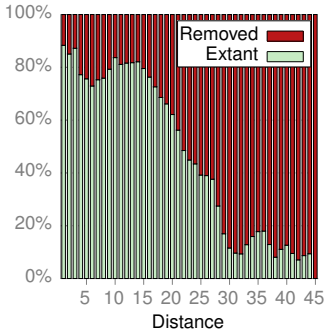

Fig. 18: Percentage users removed (spammers) versus distance from benign user.

should best reflect true clustering in the underlying social groups and interest topics. The higher clustering in the retweet graph indicates that retweet relationships are more concentrated than *following* relationships, consistent with our hypothesis of higher trust.[26]

Although the *middleman*, *in*, and *out* types are all rotations of the same basic non-transitive triplet, their coefficients differ due to the non-uniform degree distribution. C_{in} is low because the majority of (in, in) edge pairs point towards a few popular users who are retweeted by many otherwise-unrelated users. The high $C_{\text{middleman}}$ and C_{out} coefficients are reflections of the same phenomenon—transitive retweeting. User a retweeting user i's retweet of user c is recorded by Twitter as a retweeting c (hence the name middleman). Often a will also retweet some of i's original content, closing the triplet. In the *out* case, node i plays the role of a instead of the middleman. Surprisingly, such transitive retweeting happens frequently ($C_{\text{middleman}} = 0.32$ and $C_{\text{out}} = 0.29$). In other words, 30% of these possible two-degree retweet relationships exist.

6.7. Summary

We have confirmed that the retweet graph is scale-free and small-world, like many social networks. Interestingly, its clustering and assortativity behavior more closely match real-world networks than typical online networks, indicating that it may better capture real-world relationships and have application as a proxy for trust. The scale-free, small-world confirmation enables the generation of random instances, e.g., using R-MAT [Chakrabarti et al. 2004], for empirical study. We use this approach in Section 8 to evaluate the use of connectivity in the retweet graph to detect spammers.

7. IMPLICATIONS FOR THE DESIGN OF DECENTRALIZED MICROBLOGGING ARCHITECTURES

The preceding sections characterized tweet behavior—total quantity, average rate, and interevent time—and the retweet graph structure. Although interesting in their own right, in this section we discuss a particular application—the implications for the design of performance-constrained, decentralized microblogging platforms like 1am [1am 2013]. In such systems, bandwidth and energy—scarce resources—must be carefully allocated to achieve some notion of fairness. We discuss implications for such allocation strategies.

[26]N.B., not all retweets (or transitive retweets) indicate trust. Some users retweet to express *disagreement* with the original tweet.

We caution that user behaviors in a decentralized system may differ from Twitter. The strong regularities seen across other online platforms suggest that significant similarities will persist, but certain aspects might be different. For example, one's retweet behavior in Twitter may be highly dependent on whom one explicitly follows, but systems like 1am remove that control. Lacking data from a real distributed microblogging system, we use the Twitter data as the best available proxy, but note that further studies on real systems will be needed to confirm these conclusions.

Power Law Participation Momentum: Most users quit after a few contributions, so greedy allocation of resources to new users is wasteful. For example, a routing scheme prioritizing messages from users with more contributions would implicitly direct bandwidth away from temporary users.[27] The known power law form of the momentum function enables the design of optimal allocation strategies. For example, consider storing old messages by distributing them across participating nodes. Nodes with more contributions are more reliable (less likely to leave the network) and thus require a lower storage replication factor. These failure probabilities can be easily modeled.

Heavy-Tailed Rate Distribution: The two-phase tweet rate distribution has implications for short-term message delivery and long-term message storage. The message generation rate may be modeled as lognormal—messages are naturally better-distributed in the network than a power law would suggest, reducing points of congestion and better balancing bandwidth use. In the long term, however, the average tweet rates follow the asymptotic power law with its much heavier tail, posing issues for archiving and retrieval of tweets. For example, sharding [Sadalage and Fowler 2012] messages across nodes by author will result in a few nodes storing and serving the majority of the archived content. The archiving system must be designed to handle the power law distribution.

Heavy-tailed Interevent Distribution: Simulation and other performance analysis approaches must use heavy-tail distributions for interevent times. Standard Poisson distributions will grossly underpredict these times, increasing simulated congestion and resulting in over-provisioned designs.

Small-World, Assortative, Clustered Retweet Graph: In a centralized platform, a single entity can moderate bad behavior, reject spammers, and ensure fair division of resources. Participants in a decentralized platform must perform these tasks without implicit trust in others. The implicit retweet graph seems to encode some information about the real-world relationships of users that can be inferred for these purposes. Higher assortativity is more indicative of a real world network than a social network and the high clustering implies that users have some commonalities around which they gather. We explore this direction in the next section, using spammer detection via connectivity in the retweet graph as an example.

8. LEVERAGING THE RETWEET GRAPH FOR SPAMMER DETECTION

Spam is a problem for many communication platforms [Thomas et al. 2011; Yang et al. 2012], but is troublesome for decentralized, censorship-resistant microblogging platforms. Twitter, as a centralized service, can decree what constitutes spam, use its full knowledge of user behavior to detect violators [Benevenuto et al. 2010; Chen et al. 2011; McCord and Chuah 2011; Song et al. 2011; Thomas et al. 2012; Wang 2010; Yang et al. 2011], and limit the creation of new accounts. However, at its root, such filtering is a form of censorship. In a censorship-less and decentralized network, filtering must

[27]We do not consider how malicious users might manipulate such schemes, but resistance to such attacks would be important for any practical protocol.

be applied individually and locally.[28] We develop a detection approach based on the structure of the retweet graph.

8.1. Approach Overview and Background

Detection approaches can focus either on individual messages (*spam detection*) or on the sender (*spammer detection*). The latter is most applicable to microblogs, because the short message lengths (less than 250 characters) make content analysis difficult [Benevenuto et al. 2010]. Spammer detection takes two forms differentiated by the default presumption. *Blacklisting* assumes that users are not spammers until proven otherwise, while *whitelisting* presumes the opposite. The former is a non-starter in registrar-less, decentralized networks [1am 2013] because blacklisted accounts are easily and cheaply replaced. Some form of whitelisting is required.

Whitelisting presents its own issue. Manual whitelisting, akin to *following* someone on Twitter, is time-consuming and prevents previously-unacquainted users from connecting. We develop a method for automatically whitelisting users based on the intuition that non-spammers will rarely retweet spammers. This approach is easily bootstrapped by whitelisting just a few friends and, in decentralized networks, requires only local information—retweets overhead from neighbors. Blacklisting is used to block previously-good accounts that have started sending spam. Becoming whitelisted requires some effort—an account must generate content that others find worthy of sharing—and thus such accounts are not quickly or easily replaced.

Many researchers have considered spam detection in Twitter [Benevenuto et al. 2010; Chen et al. 2011; McCord and Chuah 2011; Song et al. 2011; Thomas et al. 2012; Wang 2010; Yang et al. 2011]. We survey the two most relevant works here.

Benevenuto [Benevenuto et al. 2010] studied the classification performance of 60 tweet and tweeter attributes, ranging from hashtags per tweet to the ratio of followers to friends. Aside from the obvious inclusion of URLs and account age,[29] the most sensitive attributes were related to social behavior—ratio of followers to friends, number of replies to messages, etc. Noting that spammers can easily alter the content of tweets, they suggest focusing on these harder-to-manipulate attributes for detection. Their proposed classifier has a 70% true positive rate (TPR) and a 4% false positive rate (FPR).

Song, Lee, and Kim [Song et al. 2011] developed an approach based on the followers graph that is similar to our proposal for the retweet graph. In particular, they consider two metrics in the graph: *distance*—measured as the shortest path between two nodes—and *connectivity*—measured via max-flow and random walk.[30] A classifier over these attributes had 95% TPR and 4% FPR, while the inclusion of attributes like URLs per tweet improved the performance to 99% TPR and 1% FPR.

Decentralized systems might not include explicit social relationships (e.g., 1am [1am 2013]), so the followers graph cannot be used. Instead, we consider the implicit retweet graph. Intuitively, content from spammers will not be heavily retweeted, and thus they will be less connected to non-spammers in the graph, as illustrated in Figure 17. Node A is connected to non-spammer B by three edge-independent paths, the shortest of which has length two. Spammer S, on the other hand, is only connected via a single path of length three. A separates spammers by classifying nodes based on their distance from and edge-independent connectivities (i.e., max-flow with unit weight edges) to itself.

[28] Distributed filtering mechanisms can prevent spam from propagating through the network, but a local mechanism is needed for at least the first hop.

[29] Twitter actively removes spammer accounts, biasing the collected data.

[30] The random walk metric is easily manipulated, because it requires treating the unidirectional edges as bidirectional, removing the asymmetry between spammers and non-spammers. We do not consider it further.

This scheme can be incorporated as follows. Each participant in the system maintains a partial[31] list of past messages sent by oneself and others. A partial retweet graph is constructed from this dataset, with one vertex per sender and directional edges linking each retweeter to the corresponding retweetees. Denoting the participant's own vertex as the root,[32] the remaining participants are classified by two attributes: their distance from the root and the maximum flow from the root to them. The textbook algorithms for shortest path and max flow have respective time complexities of $O(|E| + |V| \log |V|)$ and $O(|E|f)$, where f is the max flow. Both use linear space. Users classified as non-spammers are whitelisted.

This approach presents two bootstrapping problems. How does a new user with no recorded history construct a retweet graph? How are messages from a new user who has never been retweeted ever seen? For the first question, a user can copy the tweet history from a trusted friend or bootstrap by explicitly whitelisting his friends. For the second, the user can ask his friends to whitelist him, so they can then see and retweet his messages, linking him to the graph. We also anticipate that some (particularly bored) users will choose to view all incoming tweets, retweeting some that are not spam.

The following sections analyze the performance of this classification procedure on our 10% sample of the retweet graph and synthetic graphs for parameter sweeps.

8.2. Performance on the Twitter Retweet Graph

We first consider the performance on our 10% sample of the retweet graph. This sample is problematic because most of the paths between non-spammers are not included (90% of edges are missing), but is sufficient to show that the hypothesized differences exist.

We randomly chose 100 source–destination pairs of users for whom distances in the retweet graph ranged from 1 to 45, for 4500 pairs in total. We obtained ground truth classification for these 9000 users by querying the Twitter API to determine if the account had been removed in the 18 months following the initial collection. Twitter actively seeks out and bans spammers, so the majority of the spammers will have been removed. Some non-spammers will have also deleted their own accounts, so we refer to these categories as *removed* and *extant*. We believe that most removed users were spammers [Thomas et al. 2011].

We consider only the pairs whose source node is extant, because we are not interested in how well spammers can detect other spammers. Figure 18 shows the percentage of destination nodes in each category by the distance from their sources. Clearly, distance in the retweet graph is correlated with being a spammer. A classifier over this attribute alone achieves a TPR of 75% with an FPR of 25%.

The second attribute, connectivity, shows no correlation in the 10% sample graph because the majority of edges are missing. Most pairs with between one and ten independent paths in the original graph have only one or no paths in the sampled graph, making it impossible to distinguish a non-spam node linked by ten paths from a spam node linked by one. Instead, we turn to synthetic retweet graphs to study the performance of the combined classifier.

8.3. Performance on Synthetic Retweet Graphs

The analysis in Section 6 showed that the retweet graph is scale-free and small-world, enabling the generation of synthetic retweet graphs using R-MAT (*R*ecursive Matrix), an algorithm designed to generate a variety of such networks [Chakrabarti et al. 2004].

[31] Only some messages sent by others will be heard. E.g., in 1am [1am 2013], only messages broadcast in the vicinity of the node will be heard and included.

[32] Trusting one's own vertex as non-spammer breaks the otherwise problematic symmetry between the non-spammer and spammer portions of the graph.

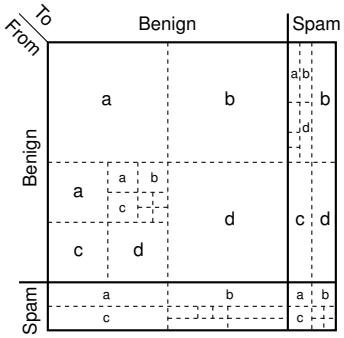

Fig. 19: Illustration of the modified R-MAT algorithm for generating synthetic retweet graphs.

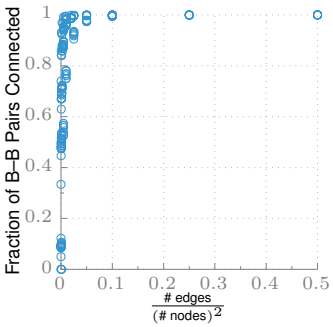

Fig. 20: Connectivity of benign pairs as a function of benign edge density.

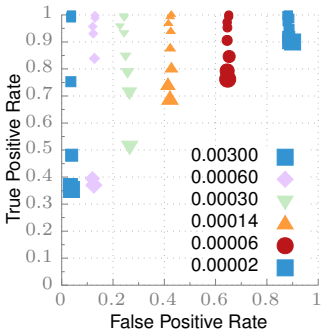

Fig. 21: J48 classifier performance over the distance and connectivity attributes.

Although metrics like assortativity and clustering are not directly controllable—R-MAT cannot capture the differences between the followers and retweet graphs[33]—it is sufficient for our purposes as we depend only on the connectivity implied by the small-world structure and the limited number of incoming edges to spammer nodes.

R-MAT produces scale-free, small-world graphs by treating edge assignment in the adjacency matrix as a two-dimensional binomial cascade. We modify the procedure to generate relatively fewer edges from benign to spammer nodes (B–S) than the other possibilities (B–B, S–S, S–B), modeling the notion that non-spammers rarely retweet spammers.

The modified R-MAT process is illustrated in Figure 19. We desire a graph with some number of benign and spammer nodes, some number of non-B–S edges, and a relatively smaller number of B–S edges. The adjacency matrix is divided into four quadrants and the edges split among the B–B, S–S, and S–B quadrants in proportion to their areas. Relatively fewer are assigned to the B–S quadrant. Within each quadrant, R-MAT is used to place the edges. For each edge, the sub-quadrant in which to place the edge is chosen according to probabilities a, b, c, and d ($a+b+c+d=1$). The process recurses until a single cell is selected for the edge.

The parameters a, b, c, and d are obtained via AutoMAT-fast [Chakrabarti et al. 2004], i.e., fitting the degree distribution of the retweet graph to that of the model. The R-MAT process is essentially a two-dimensional binomial cascade, with the out-edges assigned to the upper and lower halves with probabilities $p \triangleq a+b$ and $1-p$ and the in-edges assigned to the left and right halves with probabilities $q \triangleq a+c$ and $1-q$. Letting $N=2^n$ be the number of nodes and E the number of edges to assign, then the expected number of nodes c_k with out-degree k is $c_k = \sum_{i=0}^{n} \binom{n}{i} B\left(k; E, p^{n-i}(1-p)^i\right)$, where $B(k;a,b)$ is the mass function of the binomial distribution $B(a,b)$. The in-edge distribution is computed similarly. Fitting to the retweet graph, we obtain $a=0.52$, $b=0.18$, $c=0.17$, and $d=0.13$.

We fix the fraction of spam nodes at 10% and assume that spammer retweet behavior mimics that of benign nodes retweeting each other. Spammers could alter their behavior to gain more retweets. This is an example application, so we do not further consider such threats.

[33]Unfortunately, this prevents us from comparing retweet-based with follower-based spam detection. A full sample of the retweet graph would be needed.

The performance of the classifier is primarily affected by two metrics—the fraction of possible B–B edges that are present and the number of B–S edges per spammer vertex—so we conduct parameter sweeps of these values.

If the B–B edge density is too low, many benign pairs will not be connected and the false positive rate will be high. Figure 20 plots this density against the fraction of benign pairs that are connected for a variety of network sizes. Above 5%, most pairs are connected and above 10%, essentially all pairs are connected. We expect the number of edges in a retweet graph to (above some point) grow linearly in the number of users, so this relationship places a limit on the network size for which the technique is usable. For larger networks (e.g., the world population), the technique will only work within clusters for which the edge density is high enough—users outside of one's own cluster will be identified as spammers. For example, the average out-degree of Twitter, 75, would support 25 000 participants. However, social relationships are clustered, so this limitation should rarely be an issue in practice.[34] In a network like 1am [1am 2013], the effective community size is already limited by geography.

Figure 21 shows the classification performance. We use the J48 decision tree classifier over both the distance and connectivity (max-flow) attributes, using 10-fold cross validation. We sweep both the benign edge density (marker symbol and color) from 0.0002 to 0.003 and the number of B–S edges per spammer (marker size) from 0.01 to 1. To reduce clutter, a *single* point[35] from each resulting ROC curve is plotted. Two trends are immediately clear. Decreasing the benign edge density increases the FPR, but an FPR below 5% requires just a 0.3% edge density. Increasing the B–S rate (number of B–S edges per spammer node) decreases the true positive rate. If less than one-tenth of spammers are retweeted by benign nodes, the TPR is universally above 98%. The sensitivity to B–S rate increases with edge density because the spammer nodes are more interconnected (we hold the S–B and S–S densities equal to the B–B density).

In summary, inter-node distance in the retweet graph is highly correlated with being a spammer (Figure 18), enabling detection. Simulations on the synthetic graphs show that inter-node distance and inter-node max flow can identify spammers with greater than 98% TPR and less than 5% FPR when fewer than one-tenth of spammers are retweeted and at least 0.3% of possible edges between benign nodes are present. For a community-sized network of 25000 participants, this implies an average node degree of 75, i.e., that of the Twitter retweet graph. For larger networks, the classification works best within smaller sub-clusters where the edge density is higher.

9. CONCLUSION

We have presented an initial characterization of aggregate user behavior, describing the distributions of lifetime contributions, tweet rates, and inter-tweet durations. These behaviors are thought to be common across communication platforms, but our results differ from prior analysis, suggesting future study to determine the true extent of the similarities. Our retweet graph analysis revealed structural differences from the followers graph that are more consistent with real world social networks. Explaining the underlying causes of the observed differences is an open problem. We conjecture that retweets more closely mirror real-world relationships and trust. Finally, we developed a method for detecting spammers via their low connectivity in the retweet graph.

[34]This limitation prevents the discovery of content from outside of one's own group, which is possible with centralized Twitter today. Content can still traverse two groups if seen and retweeted by a member of both.
[35]The selected points are generally near the knees of the curves, but within a class are intentionally chosen to have similar FPRs.

REFERENCES

2013. 1am: Censorship-Resistant Microblogging. (2013). http://1am-networks.org

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (Oct. 1999), 590–512.

Albert-László Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications* 311, 3–4 (Aug. 2002), 590–614.

Albert-László Barabási and Joao Gama Oliveira. 2005. Human dynamics: Darwin and Einstein correspondence patterns. *Nature* 437, 7063 (Oct. 2005), 1251.

Christian Bauckhage, Kristian Kersting, and Bashir Rastegarpanah. 2013. The Weibull as a model of shortest path distributions in random networks. In *Proc. Wkshp. Mining and Learning with Graphs*. 1–6.

Fabrício Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida. 2010. Detecting spammers on Twitter. In *Proc. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf.* 1–9.

Catherine A. Bliss, Isabel M. Kloumann, Kameron Decker Harrison, Christopher M. Danforth, and Peter Sheridan Dodds. 2012. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *J. Computational Science* 3 (Sept. 2012), 388–397.

Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. 2003. Directed scale-free graphs. In *Proc. Symp. Discrete Algorithms*. 132–139.

Sean Borman. 2009. The Expectation Maximization algorithm: A short tutorial. (Jan. 2009). http://www.seanborman.com/publications/EM_algorithm.pdf

Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. 2005. Statistical analysis of a telephone call center. *J. American Statistical Association* 100, 469 (2005), 36–50.

Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-L'aszló Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *J. of Physics A: Mathematical and Theoretical* 41, 22 (June 2008), 224015.

Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. 2009. A measurement-driven analysis of information propagation in the Flickr social network. In *Proc. Int. World Wide Web Conf.* 721–730.

Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. 2004. R-MAT: A recursive model for graph mining. In *Proc. Int. Conf. Data Mining*. 442–446.

Xiaoling Chen, R. Chandramouli, and K.P. Subbalakshmi. 2011. Scam detection in Twitter. In *Proc. Text Mining Wkshp.* 1–10.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Rev.* 51, 4 (2009), 661–703.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statistical Society, Series B* 39, 1 (1977), 1–38.

S. N. Dorogovtsev and J. F. F. Mendes. 2000. Scaling behavior of developing and decaying networks. *Europhysics Ltrs.* 52 (Oct. 2000), 33–39.

S. N. Dorogovtsev and J. F. F. Mendes. 2001. Language as an evolving word web. *Proc. Royal Society London B* 268, 1485 (Dec. 2001), 2603–2606.

S. N. Dorogovtsev and J. F. F. Mendes. 2002. Evolution of networks. *Advances in Physics* 51, 4 (June 2002), 1079–1187.

Nick Duffield, Carsten Lund, and Mikkel Thorup. 2005. Estimating flow distributions from sampled flow statistics. *IEEE Trans. Networking* 13, 5 (Oct. 2005), 933–946.

Giorgio Fagiolo. 2007. Clustering in complex directed networks. *APS Physical Review E* 76 (Aug. 2007), 026107:1–8.

Jacob G. Foster, David V. Foster, Peter Grassberger, and Maya Paczuski. 2010. Edge direction and the structure of networks. *Proc. National Academy of Sciences of the United States of America* 107, 24 (June 2010), 10815–10820.

Miguel Freitas. 2013. Twister: Peer-to-peer microblogging. (2013). http://twister.net.co/

Maksym Gabielkov and Arnaud Legout. 2012. The complete picture of the Twitter social graph. In *Proc. Int. Conf. Emerging Networking Experiments and Technologies Student Wkshp.* 19–20.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the Twitterers—predicting information cascades in microblogs. In *Proc. Wkshp. Online Social Networks*.

Saptarshi Ghosh, Ajitesh Srivastava, and Niloy Ganguly. 2012. Effects of a soft cut-off on node-degree in the Twitter social network. *Computer Communications* 35, 7 (April 2012), 784–795.

Kwang-Il Goh and Albert-László Barabási. 2008. Burstiness and memory in complex systems. *Europhysics Ltrs.* 81, 4 (Feb. 2008), 48002.

Leo A. Goodman. 1961. Snowball sampling. *Annals Mathematical Statistics* 32, 1 (March 1961), 148–170.

Uli Harder and Maya Paczuski. 2006. Correlated dynamics in human printing behavior. *Physica A: Statistical Mechanics and its Applications* 361, 1 (Feb. 2006), 329–336.

Hai-Bo Hu and Xiao-Fan Wong. 2009. Disassortative mixing in online social networks. *Europhysics Ltrs.* 86, 1 (April 2009), 18003:1–6.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Crowdsourcing, attention and productivity. *J. Information Science* 35, 6 (Dec. 2009), 758–765.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: Understanding microblogging usage and communitites. In *Proc. Wkshp. Web Mining and Social Network Analysis*. 56–65.

Normal L. Johnson, Adrienne W. Kemp, and Samuel Kotz. 2005. *Univariate Discrete Distributions* (3 ed.). John Wiley & Sons, Inc., sec. 1.2.13.

Marcus Kaiser. 2008. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New J. Physics* 10, 8 (Aug. 2008), 083042:1–12.

Maurice George Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1–2 (June 1938), 81–93.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2006. Structure and evolution of online social networks. In *Proc. Int. Conf. Knowledge Discovery and Data Mining*. 611–617.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proc. Int. World Wide Web Conf.* 591–600. http://an.kaist.ac.kr/traces/WWW2010.html

Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. 2006. Statistical properties of sampled networks. *APS Physical Review E* 73, 1 (Jan. 2006), 016102:1–7.

Jure Leskovec and Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *Proc. Int. World Wide Web Conf.* 915–924.

Nelly Litvak and Remco van der Hofstad. 2013. Uncovering disassortativity in large scale-free networks. *APS Physical Review E* 87, 2 (Feb. 2013), 022801:1–7.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Danah Boyd. 2011. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *Int. J. Communication* 5 (2011), 1375–1405.

Alfred J. Lotka. 1926. The frequency distribution of scientific productivity. *J. Washington Academy of Sciences* 16, 12 (1926), 317–324.

M. McCord and M. Chuah. 2011. Spam detection on Twitter using traditional classifiers. In *Proc. Int. Conf. Automatic and Trusted Computing*. 175–186.

Geoffrey J. McLachlan and Thriyambakam Krishnan. 2008. *The EM Algorithm and Extensions* (2 ed.). John Wiley & Sons.

Stanley Milgram. 1967. The small-world problem. *Psychology Today* 1, 1 (May 1967), 61–67.

Staša Milojević. 2010. Power-law distributions in information science—making the case for logarithmic binning. *J. American Society for Information Science and Technology* 61, 12 (Dec. 2010), 2417–2425.

Toshio Nakagawa and Shunji Osaki. 1975. The discrete Weibull distribution. *IEEE Trans. Reliability* R-24, 5 (Dec. 1975), 300–301.

M. E. J. Newman. 2002. Assortative mixing in networks. *Physical Review Ltrs.* 89, 20 (Nov. 2002), 208701:1–4.

Christopher R. Palmer, Georgos Siganos, Michalis Faloutsos, Christos Faloutsos, and Phillip B. Gibbons. 2001. The connectivity and fault-tolerance of the Internet topology. In *Proc. Wkshp. Network-Related Data Management*. 1–6.

William J. Reed and Murray Jorgensen. 2004. The double Pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics - Theory and Methods* 33, 8 (April 2004), 1733–1753.

Pramod J. Sadalage and Martin Fowler. 2012. *NoSQL Distilled: A Brief Guide to the Emerging World of Polygot Persistence* (1 ed.). Addison-Wesley Professional.

Daniel R. Sandler and Dan S. Wallach. 2009. Birds of a FETHR: Open, decentralized micropublishing. In *Proc. Int. Wkshp. Peer-to-Peer Systems*. 1–6.

Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskovec. 2008. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proc. Int. Conf. Knowledge Discovery and Data Mining*. 596–604.

S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, and M. Paczuski. 2012. Sampling properties of directed networks. *APS Physical Review E* 86, 4 (Oct. 2012), 046104:1–12.

Jonghyuk Song, Sangho Lee, and Jong Kim. 2011. Spam filtering in Twitter using sender–receiver relationship. In *Proc. Int. Symp. Recent Advances in Intrusion Detection*. 301–317.

Pierre St Juste, David Wolinsky, P. Oscar Boykin, and Renato J. Figueiredo. 2011. Litter: A lightweight peer-to-peer microblogging service. In *Proc. Int. Conf. Privacy, Security, Risk and Trust*. 900–903.

William E. Stein and Ronald Dattero. 1984. A new discrete Weibull distribution. *IEEE Trans. Reliability* R-33, 2 (June 1984), 196–197.

Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc. National Academy of Sciences of the United States of America* 102, 12 (March 2005), 4221–4224.

Bongwon Suh, Lichan Hong, Petr Pirolli, and Ed H. Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proc. Int. Conf. Social Computing*. 177–184.

Ole Tange. 2011. GNU Parallel—the command-line power tool. *;login: The USENIX Magazine* 36, 1 (Feb. 2011), 42–47. http://www.gnu.org/s/parallel

Abraham Ronel Martínez Teutle. 2010. Twitter: Network properties analysis. In *Proc. Int. Conf. Electronics, Communications, and Computer*. 180–186.

Kurt Thomas, Chris Grier, and Vern Paxson. 2012. Adapting social spam infrastructure for political censorship. In *Proc. Wkshp. Large-Scale Exploits and Emergent Threats*.

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: An analysis of Twitter spam. In *Proc. Internet Measurement Conf.* 243–256.

Jeffrey Travers and Stanley Milgram. 1969. An experimental study of the small world problem. *Sociometry* 32, 4 (Dec. 1969), 425–443.

Alex Hai Wang. 2010. Don't follow me: Spam detection in Twitter. In *Proc. Int. Conf. Security and Cryptography*. 1–10.

Audrey Watters. 2011. How recent changes to Twitter's terms of service might hurt academic research. (March 2011). http://webcitation.org/6MgAFaaMi http://readwrite.com/2011/03/03/how_recent_changes_to_twitters_terms_of_service_mi.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (June 1998), 440–442.

Dennis M. Wilkinson. 2008. Strong regularities in online peer production. In *Proc. Conf. Electronic Commerce*. 302–309.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on Twitter. In *Proc. Int. World Wide Web Conf.* 705–714.

Tianyin Xu, Yang Chen, Jin Zhao, and Xiaoming Fu. 2010. Cuckoo: Towards decentralized, socio-aware online microblogging services and data measurements. In *Proc. HotPlanet Wkshp.* 1–6.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammer's social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter. In *Proc. Int. World Wide Web Conf.*

Chao Yang, Robert Chandler Harkreader, and Guofei Gu. 2011. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In *Proc. Int. Symp. Recent Advances in Intrusion Detection*. 318–337.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proc. Int. Conf. Web Search and Data Mining*. 177–186.

Haifeng Yu, P.B. Gibbons, M. Kaminsky, and Feng Xiao. 2008a. SybilLimit: A near-optimal social network defense against Sybil attacks. In *Proc. Symp. Security and Privacy*. 3–17.

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. 2008b. SybilGuard: Defending against Sybil attacks via social networks. *IEEE Trans. Networking* 16, 3 (June 2008), 576–589.