Statistic laws governing rumor popularity.

Researchers have noticed for decades that many measured data retrieved from biological and social sciences can be described by log-normal distribution[23,24] and power-law distribution[25]. In this case study, we estimate the log-normal and power-law models of our collected rumor popularity data, including both news dataset and Twitter dataset. We utilize *powerlaw* package in python to estimate the fitted models[26]. We perform a statistical hypothesis test analysis as follows: (i) We estimate the parameters of fitted models via *powerlaw*. For example, for the fitted power law model, the estimated parameter α indicates the index of power law's probability distribution: $p(x) \propto x^{\alpha}$; and the estimated x_{min} indicates the optimal start point of the power law fit (tail). (ii) After parameter estimation, we calculate the goodness-of-fit between the popularity data and the power-law (and log-normal). Specifically, we calculate p_{KS}, a plausibility value, based on measurement of the "distance" between the distribution of the empirical data and the hypothesized model. The distance D is estimated by *powerlaw* when we fit the data, which is the "Kolmogorov-Smirnov (KS)" statistic. Next, we generate a large number of power-law (and log-normal) synthetic data with the estimated parameters and we fit the synthetic data using *powerlaw*. After fitting the synthetic data, we get the distance of synthetic data and the hypothesized power-law model (and log-normal model fitted by the synthetic data), noted as D_{syn}. Then we repeat this procedure by generating 100 sets of synthetic data with 100 D_{syn}'s. Finally we calculate p_{KS} as the percentage of $D_{syn} > D$. If p_{KS} is greater than 0.1, the power-law (or log-normal) is a plausible hypothesis for the data[27]. (iii) After calculating the p_{KS}, we compare hypotheses, power-law and log-normal, via a likelihood ratio test provided in *powerlaw*, e.g., $R, p = distribution\_compare('lognormal', 'powerlaw')$, where R is the log-likelihood ratio between the two candidate distributions. If $R > 0$, then the data are more likely to follow the first distribution, otherwise the data are more likely to obey the second distribution. p is the significance value for that direction. The favored distribution is a strong fit if $p > 0.05$. Following the hypothesis test, our estimated results of rumor popularity data are shown in Figure 10.

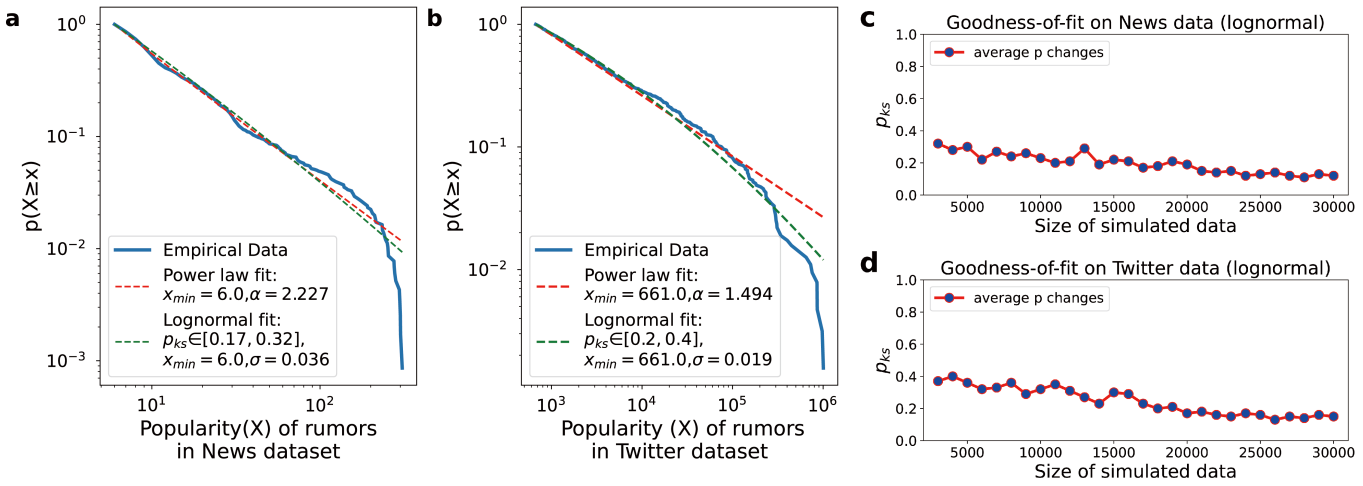

Figure 10. Comparison between log-normal and power-law fit of twitter rumors and the goodness of fit test on log-normal distribution. The p_{KS} on log-normal fit of Twitter data falls in $[0.20, 0.40]$. The result indicates the original data matches log-normal distribution well. The p_{KS} of news data falls in $[0.17, 0.32]$, which also means the original data matches log-normal distribution well.

The distribution fitting comparison between power law and log-normal for News and twitter data are shown in Figure 10. The estimation and the resulting parameters for the power law and the log-normal are shown in Figure 10a-b. The likelihood ratio test results for news and Twitter data are $R, p = (-3.0001, 0.0027)$ and $R, p = (-3.1880, 0.0014)$, respectively. The negative R values indicate that the data are more likely to follow the second distribution, i.e., log-normal. The goodness-of-fit tests are therefore performed for log-normal hypothesis and the results are shown in 10c-d. With p_{KS} values greater than 0.1 in all experiments, we conclude that the rumor popularity data (from both news and Twitter dataset) are indistinguishable from identically and independently drawn samples from the log-normal distribution.