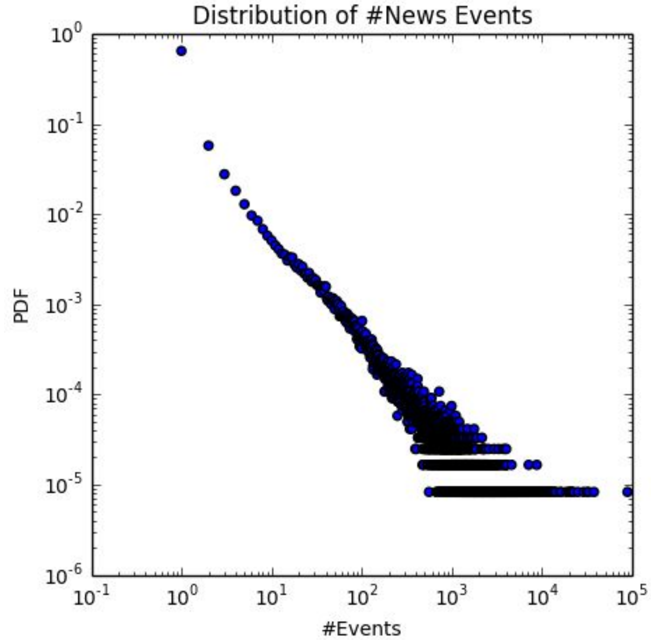
This is an example showing how we can validate the hypothesis that a distribution follows the power-law. For more details from the theoretical aspects, please refer to the paper "Power-law distributions in empirical data." by Clauset et al. or our slides attached in this github repo.

Power-Law Distribution

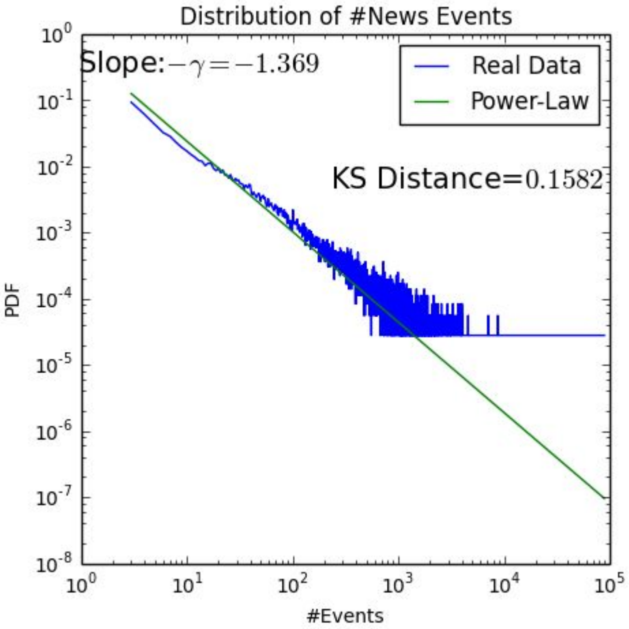
Xiaoyan Lu

2017 Spring

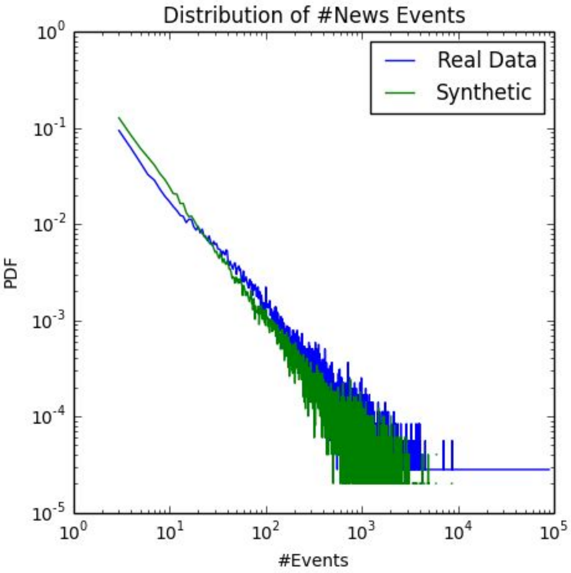
Using Google BigQuery, we count the number of news events reported by each news site in the last two years in the GDelt dataset. The scatter plot below shows the distribution the number of news events reported by each new site. Note that the each bin has the width 50. In other words, the news sites which have reported at least 50*k and no more than 50*(k+1) events are arranged in the k-th bin.



Then, we can fit the data into a Power-law model. The exponent “gamma” is computed by MLE estimator and the Kolmogorov–Smirnov distance is computed based on the CDF of the real distribution and the CDF of the Power-law model. The minimum “degree” where the power-law distribution starts is 2 here. We have found that a larger minimum “degree” can result in a smaller KS distance, but it also means more data gets dropped. So we simply take 2 as the lower bound.



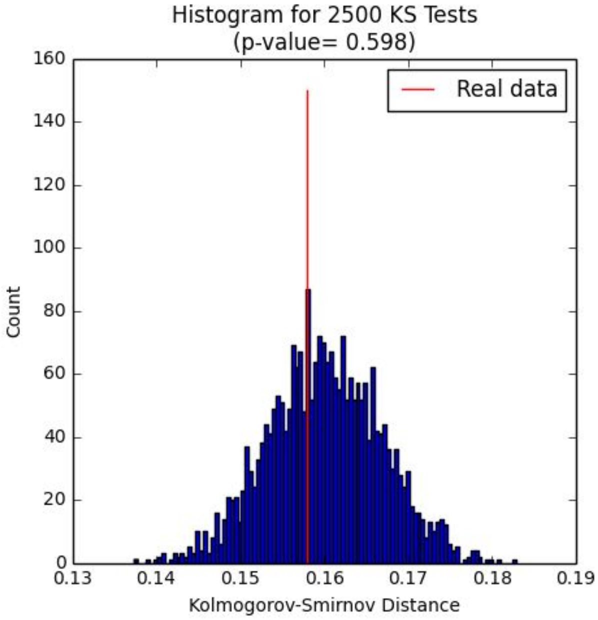
In order to compute the p-value, we generate 2,500 synthetic distributions according to the obtained power law and compute KS distance for each of the hypothetical sequences. A sample synthetic sequence is plotted as follows.



The goodness of fit is measured by

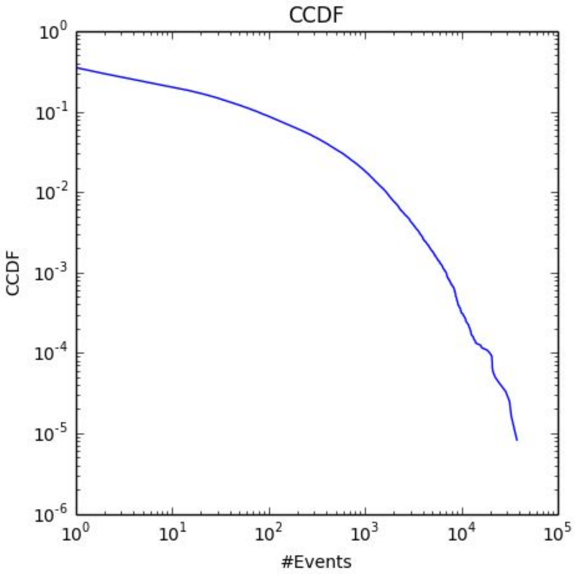
$$p = \int_{D^{real}}^{\infty} p(D^{synthetic})\, dD^{synthetic}$$

where $D_{synthetic}$ is the KS distance of the synthetic data and D_{real} is the KS distance of the real data, 0.1582 in our case. The obtained p-value is 0.598 which is much larger than 10%. It implies the Power-law model is a plausible fit to the data.

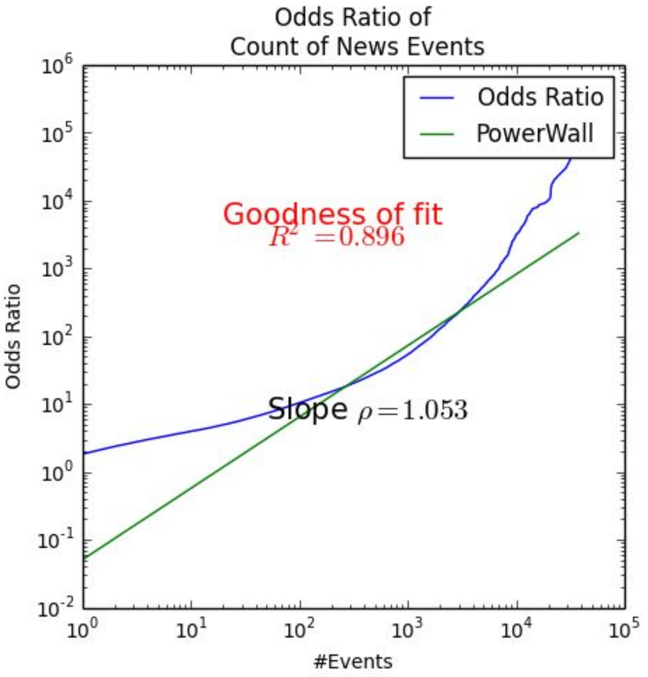


PowerWall Model

We also consider the PowerWall model proposed by Devineni, Pravallika, et al.. The CCDF of the distribution of the #events is plotted below.



Since the Odds Ratio of a log-logistic distribution is linear on a double logarithmic plot, we plot the Odds Ratio in log-log scale and do linear regression.



The goodness of fit is computed by the coefficient of determination, also known as R2. The R^2 = 0.896 which is smaller than the values found in the Facebook wallposts data.

Reference

Devineni, Pravallika, et al. "If walls could talk: Patterns and anomalies in Facebook wallposts." *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review* 51.4 (2009): 661-703.