Harder, better, faster, stronger cascades – or simply larger?

Jonas L. Juul[1] and Johan Ugander[2]

[1] *Center for Applied Mathematics, Cornell University, Ithaca, New York 14853, USA*
[2] *Management Science & Engineering, Stanford University, Stanford, CA, 94305 USA*

Does false news and true news diffuse differently online? How about visual vs. written content? To answer such questions rigorously, scholars study diffusion cascades — collections of timestamped, rooted, directed trees. By using statistical network analysis to study the structure of these trees for different online content, the hope is to understand how the content spreads online, and how diffusion of different content types differ. Here, we provide new theory and data analysis to demonstrate the importance of joint distributions of key statistical properties of diffusion cascades in any analysis [1].

Diffusion cascade structures have played key parts in a number of recent impactful papers. Perhaps the most prominent example is Vosoughi *et al.* [2], who studied a comprehensive dataset of trees for all fact-checked true and false content that spread on Twitter in the years 2006-2017. Comparing the statistics of false-news and true-news cascades, they concluded that false news spreads "significantly farther, faster, deeper, and more broadly" than the truth. Fig. 1A-D show the differences in size, depth, breadth and propagation speed that lead Vosoughi et al. to this conclusion. However, cascade size, depth, and max breadth are not necessarily statistically independent: Doubling the size of a cascade, the cascade might increase in depth or maximal breadth too. Similarly, propagation speed might correlate with cascade size: larger cascades might reach 1 000 adoptions faster than smaller cascades would.

To resolve whether the reported differences in true-news and false-news cascade structures are simply due to tree size differences, we control for the effect of cascade size. We match sizes of false-news to true-news cascades, creating two new data sets with identical size distributions. Repeating Vosoughi *et al.*'s analysis on the size-matched data sets in Fig. 1F-I, we see that cascade structures and propagation speed are indistinguishable for true and false news when cascade sizes are identical.

What can the indistinguishability of cascade structure for size-matched data sets tell us about the spreading mechanisms for true and false news? We prove that a similar collapse of statistical cascade differences take place when comparing sets of cascades created under the same branching process with different parameter choices. Fig. 1E,J demonstrate such a collapse for two datasets of trees created by simulating an SIR model on infinite complete graphs with different choices for the infectiousness parameter, R_0. Size matched datasets have indistinguishable cascade structures. Crucially, this collapse does not happen when comparing size-matched data sets of cascades that were created under different diffusion models. Interestingly, we demonstrate the absence of the collapse when comparing size-matched empirical cascades of diffusing videos, pictures, news, and petitions [3].

Analyzing cascade structures is a key tool to study how content spreads on online platforms. Alas, joint distributions of cascade structure have not been controlled for previously. We argue that carefully controlling for size and other relevant differences may help reveal whether data sets of cascades were created under similar diffusion mechanisms. False news appears to spread further than true news, but whether its diffusion is otherwise different is unclear.

---

[1] J L Juul, J Ugander, *Proceedings of the National Academy of Sciences*, 118(46), e2100786118 (2021).

[2] S Vosoughi, D Roy, S Aral, *Science* **359**, 1146–1151 (2018).

[3] S Goel, A Anderson, J Hofman, D J Watts. (2016).Management Science, 62(1), 180-196.

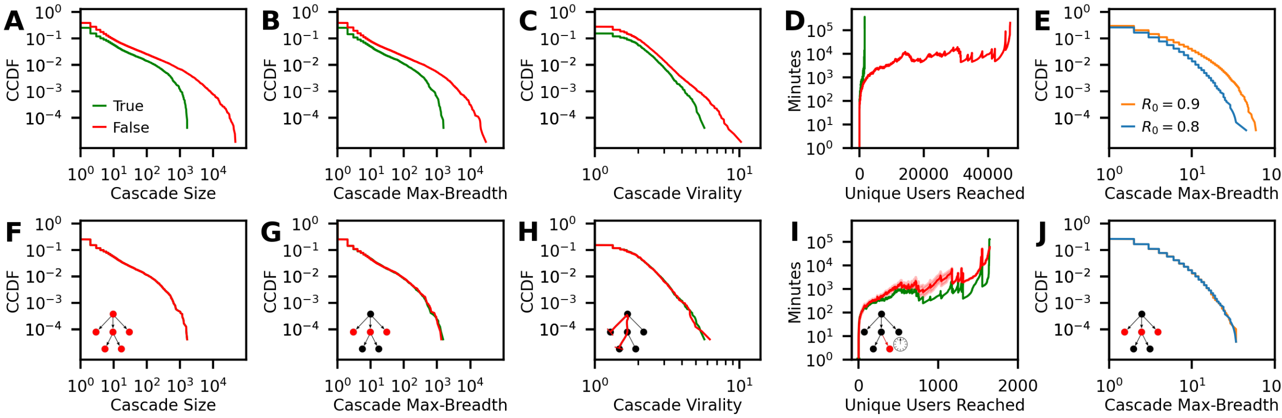

FIG. 1. **A-D** Topological and temporal statistics of false-news and true-news cascades diffusing on Twitter, as presented in [2]. Cascades in the two datasets have different size distributions (see **A**). **E** Example statistic for two SIR models with different R_0. Cascades are larger and broader for $R_0 = 0.9$. **F-J** Same analyses as the plots directly above, carried out for two subsampled datasets with identical size distributions. Distributions are indistinguishable for bottom plots. Panels, caption adapted from [1].