Evidence that coronavirus superspreading is fat-tailed

Felix Wong[a,b,c] and James J. Collins[a,b,c,d,1]

[a]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142; and [d]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115

Edited by Simon A. Levin, Princeton University, Princeton, NJ, and approved September 28, 2020 (received for review September 1, 2020)

Superspreaders, infected individuals who result in an outsized number of secondary cases, are believed to underlie a significant fraction of total SARS-CoV-2 transmission. Here, we combine empirical observations of SARS-CoV and SARS-CoV-2 transmission and extreme value statistics to show that the distribution of secondary cases is consistent with being fat-tailed, implying that large superspreading events are extremal, yet probable, occurrences. We integrate these results with interaction-based network models of disease transmission and show that superspreading, when it is fat-tailed, leads to pronounced transmission by increasing dispersion. Our findings indicate that large superspreading events should be the targets of interventions that minimize tail exposure.

COVID-19 | SARS-CoV-2 | superspreading | extreme value theory | infectious disease

Superspreading has been recognized as an important phenomenon arising from heterogeneity in individual disease transmission patterns (1). The role of superspreading as a significant source of disease transmission has been appreciated in outbreaks of measles, influenza, rubella, smallpox, Ebola, monkeypox, SARS, and SARS-CoV-2 (1, 2). A basic definition of an nth-percentile superspreading event (SSE) has been proposed to be any infected individual who infects more people than does the nth-percentile of other infected individuals (1). Hence, if the number of secondary cases is randomly distributed, then for large n, SSEs can be viewed as right-tail events. A natural language for understanding the tail events of random distributions is extreme value theory, which has been applied to contexts as diverse as insurance (3) and contagious diseases (4). Here, we apply extreme value theory to empirical data on superspreading in order to gain insight into this critical phenomenon impacting the current COVID-19 pandemic.

Results and Discussion

We view the number of secondary cases resulting directly from an index case of a disease to be a random variable, Z. We also view the individual reproductive number, ν, to be a random variable representing the expected number of secondary cases caused by an infected individual. Seminal work (1) has suggested that, for SARS-CoV, Z follows a negative binomial distribution, $Z\sim$negative binomial(R_0,k), where R_0 is the basic reproduction number, k is the dispersion parameter quantifying variation in transmission, and the mean and variance of Z are R_0 and $R_0(1+R_0/k)$, respectively. Assuming that stochastic effects in transmission are modeled by a Poisson process, ν is gamma-distributed and $1/k$ effectively measures the "flatness" of the distribution of ν. Different assumptions of the branching process can be modeled, and we focus on the foregoing assumptions for simplicity (1). For SARS-CoV, k has been estimated to be ∼0.16 (1); for SARS-CoV-2, k has been estimated to be ∼0.1 to 0.6 (2, 5). Importantly, if $Z\sim$negative binomial(R_0,k), then for $k \leq 1$, Z has an exponential tail (6). This means that the occurrence of SSEs has a probability that decreases exponentially as Z increases.

Tails are exceptionally significant in extreme value theory, where they determine how rare extreme events are, how the central limit theorem is generalized, and what distribution the scaled maxima of samples follow. We were therefore interested to determine whether the empirically observed distribution of Z for SARS-CoV and SARS-CoV-2 exhibited an exponential tail. We searched the scientific literature for global accounts of SSEs, in which single cases resulted in numbers of secondary cases greater than R_0, estimated to be ∼3 to 6 for both coronaviruses (1, 7). To broadly sample the right tail, we focused on SSEs resulting in >6 secondary cases, and as data on SSEs are sparse, perhaps due in part to a lack of data sharing, we pooled data for SARS-CoV and SARS-CoV-2. Moreover, to avoid higher-order transmission obfuscating the cases generated directly by the index case, we ruled out SSEs where a single infected individual led to a cluster of subsequent infections, but the subsequent infections were not indicated to be secondary cases.

Curating a total of 60 SSEs in this way, we found 45 SSEs associated with SARS-CoV-2 and 15 SSEs associated with SARS-CoV (Fig. 1 *A* and *B*). An additional 14 SSEs were documented in news sources and not scientific studies, and their inclusion does not significantly change the following results, which also hold when accounting for sources of bias (below). Details of the dataset are summarized in Dataset S1.

Several striking observations emerge from the data. While the SSEs surveyed indicated secondary case numbers ranging from ∼6 in a family-spreading incident in Singapore to 187 in an apartment in Hong Kong, many SSEs exhibited significantly more secondary cases than $R_0 \approx 3$ to 6, with the conditional sample mean being 19.7 cases (Fig. 1 *A* and *B*).

We next examined the tail behavior of Z using inference tools from extreme value theory. We found that the tail of Z, as sampled by our list of SSEs, $\{Z_i\}$, was inconsistent with exponential decay. Instead, we found that the tail of Z is consistent with fat-tail behavior using three complementary methods: 1) a Zipf plot; 2) a meplot; and 3) statistical estimators of the tail index, which collectively suggest a power-law scaling of the form $\Pr(Z > t)\sim t^{-\alpha}$ for large t, with α between 1 and 2 (Fig. 1 *C*–*E* and *SI Appendix*, *Methods*). Equivalently, this observation indicates that the tails of Z—as quantified by the threshold exceedance values $\{Z_i - u | Z_i \geq u\}$—can be described by the generalized Pareto distribution, with corresponding tail index $\xi = 1/\alpha$ between 0.5 and 1. That $\xi \leq 1$ is significant, since all moments higher than $1/\xi$ diverge for a generalized Pareto distribution (3).

Our finding that the tail of Z is fat has implications not only for superspreading, but also for modeling the effects of individual variation on disease transmission. First, the fat tail of Z makes the distribution of Z inconsistent with a negative binomial distribution, and the consistency of the tail with a generalized Pareto distribution suggests that it arises from branching processes in which the time to infection, instead of ν, is gamma-distributed

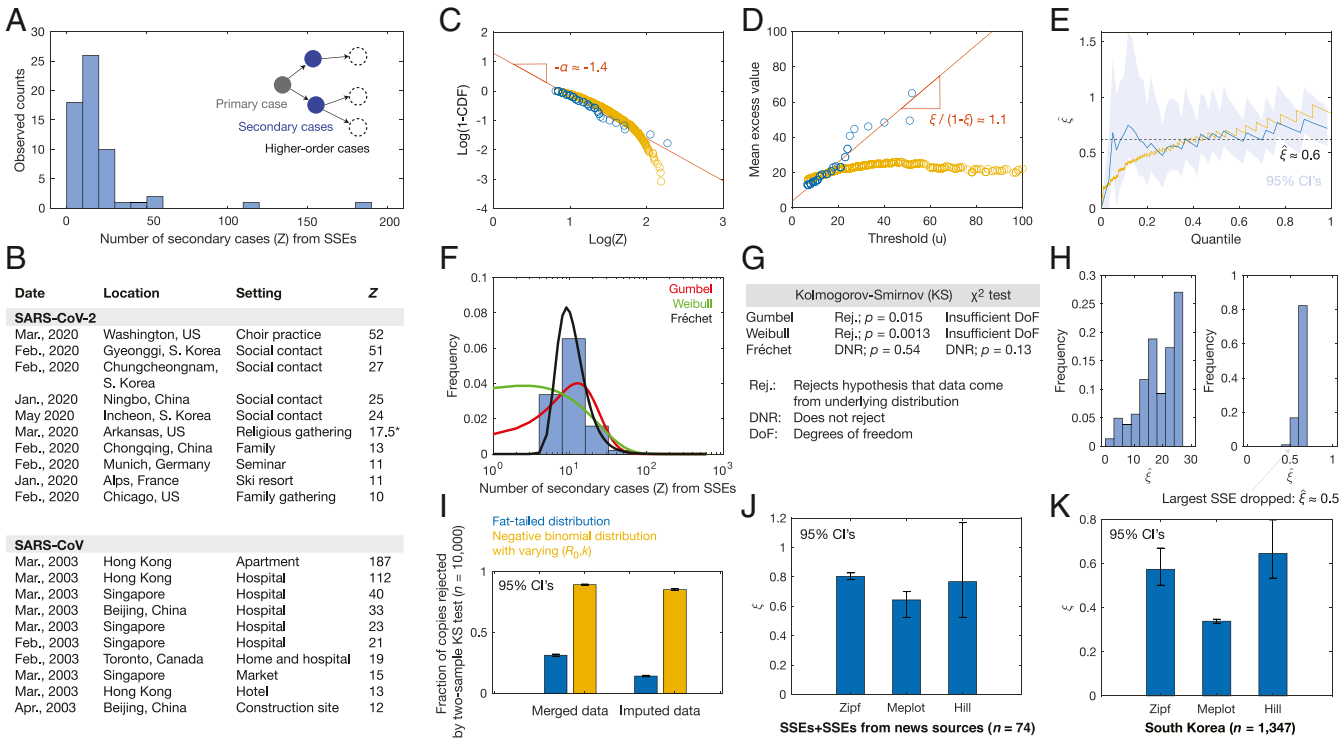
Author contributions: F.W. and J.J.C. designed research; F.W. performed research; F.W. contributed new reagents/analytic tools; F.W. and J.J.C. analyzed data; and F.W. and J.J.C. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jimjc@mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2018490117/-/DCSupplemental.

First published November 2, 2020.



Date	Location	Setting	Z
SARS-CoV-2			
Mar., 2020	Washington, US	Choir practice	52
Feb., 2020	Gyeonggi, S. Korea	Social contact	51
Feb., 2020	Chungcheongnam, S. Korea	Social contact	27
Jan., 2020	Ningbo, China	Social contact	25
May 2020	Incheon, S. Korea	Social contact	24
Mar., 2020	Arkansas, US	Religious gathering	17.5*
Feb., 2020	Chongqing, China	Family	13
Feb., 2020	Munich, Germany	Seminar	11
Jan., 2020	Alps, France	Ski resort	11
Feb., 2020	Chicago, US	Family gathering	10
SARS-CoV			
Mar., 2003	Hong Kong	Apartment	187
Mar., 2003	Hong Kong	Hospital	112
Mar., 2003	Singapore	Hospital	40
Mar., 2003	Beijing, China	Hospital	33
Mar., 2003	Singapore	Hospital	23
Feb., 2003	Singapore	Hospital	21
Feb., 2003	Toronto, Canada	Home and hospital	19
Mar., 2003	Singapore	Market	15
Mar., 2003	Hong Kong	Hotel	13
Apr., 2003	Beijing, China	Construction site	12

Fig. 1. SARS-CoV and SARS-CoV-2 SSEs correspond to fat tails. (*A*) Histogram of *Z* for 60 SSEs. (*B*) Subsample of 20 diverse SARS-CoV and SARS-CoV-2 SSEs. *See Dataset S1 for details. (*C*) Zipf plots of SSEs (blue) and 10,000 samples of a negative binomial distribution with parameters $(R_0,k) = (3,0.1)$, conditioned on $Z > 6$ (yellow). (*D*) Meplots corresponding to *C*. (*E*) Plots of $\hat{\xi}$, the Hill estimator for ξ, for the samples in *C*. (*F*) Different extreme value distribution fits to the distribution of SSEs. (*G*) One-sample Kolmogorov–Smirnov and χ^2 goodness-of-fit test results for the fits in *F*. (*H*) Robustness of results, accounting for noise (*Left*) and incomplete data (*Right*). (*I*) Inconsistency of the maxima of 10,000 samples of a negative binomial distribution (yellow) with the SSEs in *A*, accounting for variability in (R_0,k) and data merging and imputation, in contrast to the maxima of 30 samples from a fat-tailed (Fréchet) distribution (blue) with tail parameter $\alpha = 1.7$ and mean $R_0 = 3$. The numbers of samples in each case were determined so that the sample mean of maxima is equal to the sample mean from *A*. (*J*–*K*) Generality of inferred ξ to 14 additional SSEs from news sources (*J*) and a dataset of 1,347 secondary cases arising from 5,165 primary cases in South Korea (*K*) (Dataset S2).

(so that the tails of Z correspond to an exponential-gamma mixture); this prediction is consistent with studies that have fitted serial intervals to gamma distributions (8, 9). Second, since the second moment of Z diverges if $\alpha < 2$, the occurrence of SSEs suggests that measuring variances of empirical samples of Z can be misleading. Third, fat-tailed distributions generate extreme risk, and superspreading should be mitigated by measures that reduce tail events instead of focusing on the bulk of the distribution.

A complementary way in which we may interpret superspreading is by assuming that SSEs arise not only as right-tail samples of Z, but also as the maxima of many samples of the entire distribution of Z. The consistency of this viewpoint with the definition of SSEs as right-tail samples of Z is given by an important theorem in extreme value theory relating threshold exceedances to extreme value distributions (3). Indeed, SSEs often represent the maxima of values of Z observed in transmission clusters. In this case, the Fisher–Tippett–Gnedenko theorem asserts that distributions of the maximum of large numbers of samples converge to either the Gumbel, Fréchet, or Weibull distributions if the tails of the underlying distribution are exponentially decaying, fat, or thin (faster-than-exponential) and finite, respectively. Supporting the view of SSEs as maxima of ensembles of spreading events, we found that the distribution of observed SSEs was consistent with the Fréchet distribution but inconsistent with the Gumbel and Weibull distributions, as measured by maximum-likelihood fitting and one-sample Kolmogorov–Smirnov and χ^2 goodness-of-fit tests at the 5% significance level (Fig. 1 *F* and *G* and *SI Appendix*, *Methods*).

We next verified that our results were robust to noisy and incomplete data (4). To account for noise, we generated 10,000 copies of the data, where each copy involved multiplying the original data by uniform random variables in [0.5,1.5]—a range that we anticipate to accommodate errors in testing and reporting—and recomputed $\hat{\xi}$ according to the Hill estimator (*SI Appendix*, *Methods*). To account for incomplete data, a random number of observations between 1 and 10 was randomly removed, according to uniform distributions, for 10,000 copies of the data, and $\hat{\xi}$ was recomputed. The variation in $\hat{\xi}$ is summarized in Fig. 1*H*. Notably, we observed that $\hat{\xi}$ was always greater than 0.5, so that the second and higher moments of Z diverge.

In a complementary analysis, we tested for sources of bias in the data, which could arise from variations in testing and reporting. As null models, we tested whether the data could be consistent with the maxima of samples from a negative binomial distribution with (R_0,k) randomly sampled in $[0,6] \times [0,1]$ and in which up to 40% of entries were merged or imputed by the mean. Statistical tests of 10,000 copies of simulated data indicated that these sources of variation cannot explain the observed SSEs, which instead favor an underlying fat-tailed distribution despite this variation (Fig. 1*I*). Moreover, we repeated our analyses after adding 14 SSEs from news sources and for a contact-tracing dataset of 1,347 secondary cases arising from 5,165 cases in South Korea (10) (Dataset S2). We found that both datasets exhibited fat-tailed behavior, with inferred tail indices ($\xi \approx 0.3$ to 0.8) quantitatively similar to those found above (Fig. 1 *J* and *K*).

Combining these results with modeling can be timely for informing interventions in the current pandemic. As a proof of concept, we considered a network model of transmission which fine-grains an SEIR model (Fig. 2*A*). Here, 1,000 individuals

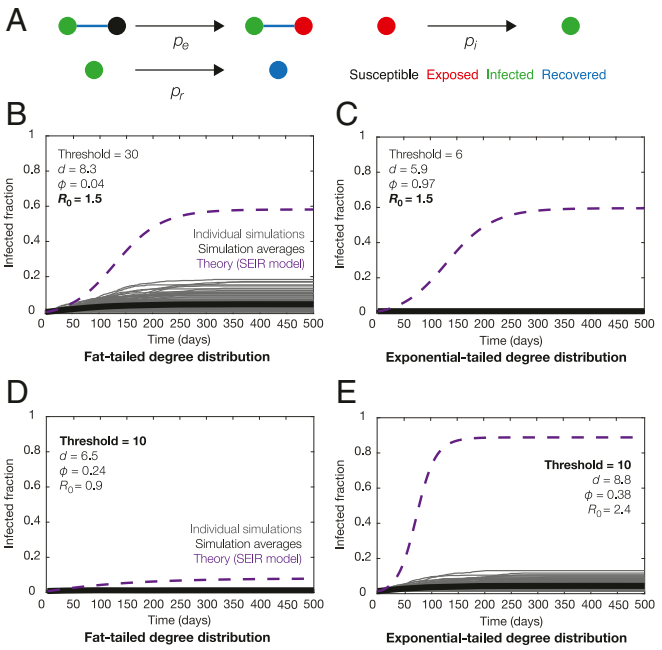


Fig. 2. Forward modeling of intervention strategies. (*A*) State transitions in a fine-grained network model of disease transmission. (*B–E*) Predicted total infected fraction for an intervention strategy that isolates a fraction φ of all individuals, namely those with degree greater than the threshold number, and yielding decreased mean connectivity of d and effective basic reproduction number of R_0. Here, R_0 depends on the coefficient of variation of the degree distribution, as detailed in Dataset S3. Trajectories from 100 simulations for BA random graphs (*B* and *D*) and WS random graphs (*C* and *E*) and their averages are shown, compared to the theoretical predictions for a well-mixed model.

(nodes) each transition between being susceptible (S), exposed (E), infected (I), and recovered or dead (R) with rates $S \xrightarrow{\beta SE} E$, $E \xrightarrow{\delta E} I$, and $I \xrightarrow{\gamma I} R$, as detailed further in Dataset S3, and rates were chosen with $R_0 = 3$ and a characteristic incubation time of 5 days for SARS-CoV-2 (7). We considered two different graph models with identical mean connectivity ($m = 10$): Barabási–Albert (BA) and Watts–Strogatz (WS), which possess fat-tailed ($\alpha = 2$) and exponential-tailed degree distributions, respectively. As a simple intervention strategy, we considered node removals in which a fraction φ of all nodes is removed starting from those with largest degree. We found that, when the degree threshold for node removals was chosen to yield the same effective value of R_0 in both models, the BA model resulted in greater transmission (Fig. 2 *B* and *C*), indicating that a fat-tailed degree distribution contributes to transmission by increasing dispersion. In contrast, for the same degree threshold, we found that isolating all possible superspreaders—defined here as individuals with degree greater than 10, corresponding to the 80th percentile in the BA model and the 50th percentile in the WS model—suffices to decrease R_0 below 1 and control the pandemic for the BA, but not WS, model (Fig. 2 *D* and *E*). Intriguingly, in both models, stochastic extinction events lead to smaller infected fractions than those predicted by a well-mixed model (Fig. 2 *B–E*). These results indicate that transmission is especially pronounced when superspreading is fat-tailed and hint at more detailed models of interventions focused on tail events. We anticipate future models to consider not only heterogeneity in network interactions, but also in infectivity and susceptibility (11).

In summary, we have provided evidence that the distribution of secondary cases, Z, is fat-tailed with tail exponent $\alpha \in [1,2]$. The fat-tailed nature of Z indicates that SSEs have an outsized contribution to overall transmission and should be the targets of interventions that minimize tail exposure, for instance, by preventing large gatherings of susceptible individuals or immunizing select individuals (12). Extreme value theory offers a framework for modeling superspreaders, and we anticipate that using the tools of this theory can, as illustrated here, better allow us to understand the effects of superspreading on the ongoing pandemic.

Data Availability. All analysis code are available at GitHub, https://github.com/felixjwong/superspreaders. All study data are included in the article and *SI Appendix*.

ACKNOWLEDGMENTS. We thank the editor and the three anonymous reviewers for helpful suggestions, and Po-Yi Ho and Jie Lin for helpful comments. F.W. was supported by the James S. McDonnell Foundation.

1. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
2. D. C. Adam *et al.*, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.*, 10.1038/s41591-020-1092-0 (2020).
3. P. Embrechts, C. Klüppelberg, K. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Stochastic Modelling and Applied Probability, Springer, 1997).
4. P. Cirillo, N. N. Taleb, Tail risk of contagious diseases. *Nat. Phys.* **16**, 606–613 (2020).
5. A. Endo, S. Abbott, A. J. Kucharski, S. Funk; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
6. P. Yan, "Distribution theory, stochastic processes and infectious disease modeling" in *Mathematical Epidemiology*, F. Brauer, P. van den Driessche, J. Wu, Eds. (Lecture Notes in Mathematics, Springer, 2008).
7. Y. M. Bar-On, A. Flamholz, R. Phillips, R. Milo, SARS-CoV-2 (COVID-19) by the numbers. *eLife* **9**, e57309 (2020).
8. X. He *et al.*, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
9. J. Zhang *et al.*, Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: A descriptive and modelling study. *Lancet Infect. Dis.* **20**, 793–802 (2020).
10. J. Kim *et al.*, *DS4C:* Data Science for COVID-19 in South Korea. https://www.kaggle.com/kimjihoo/coronavirusdataset. Accessed 28 August 2020.
11. T. Britton, F. Ball, P. Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849 (2020).
12. R. Cohen, S. Havlin, D. Ben-Avraham, Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.* **91**, 247901 (2003).

PNAS
www.pnas.org

Supplementary Information for

Evidence that coronavirus superspreading is fat-tailed

Felix Wong[1,2] and James J. Collins[1,2,3,*]

[1]Institute for Medical Engineering & Science and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[2]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
[3]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA.

***Email:** jimjc@mit.edu

This PDF file includes:

Extended Methods

The Zipf plot shown in Fig.1C of the main text is a log-log plot of the survival function against the number of secondary cases, and the linearly decreasing behavior it shows suggests a power-law scaling of the form $\Pr(Z>t)\sim t^{-\alpha}$ for large t. The value of the power-law coefficient, $\alpha\approx1.45$ (95% CI: [1.38,1.51]), is greater than 1. Equivalently, this observation indicates that the tails of Z—as quantified by the threshold exceedance values $\{Z_i-u|Z_i\geq u\}$—can be described by the generalized Pareto distribution, with corresponding tail index $\xi=1/\alpha\approx0.7$ (95% CI: [0.62,0.76]). That $\xi\leq1$ is significant, since all moments higher than $1/\xi$ diverge for a generalized Pareto distribution (1).

The Zipf plot can be complemented by computing the mean excess function of Z, $e(u)=\mathrm{E}(Z-u|Z\geq u)$, which for a generalized Pareto distribution is linear in u with slope $\xi/(1-\xi)$ (1). Hence, checking for linearity in a plot of u against $e(u)$ — a mean excess plot — above some threshold u allows one to verify the existence of fat tails. We observed in a meplot that for $u>10$, $e(u)$ indeed increases approximately linearly with a slope of ~1.11 (Fig.1D; 95% CI: [1.02,1.20]; adjusted R^2: 0.91), suggesting a value of $\xi\approx0.5$, which is qualitatively consistent with the Zipf plot of Fig.1C of the main text.

The Hill estimator of the tail index ξ is

$$\hat{\xi}(k)=\frac{1}{k}\sum_{i=1}^{k}\log\left(Z_{i,n}/Z_{k,n}\right),$$

where $2\leq k\leq n$ and $Z_{n,n}\leq Z_{n-1,n}\leq\ldots\leq Z_{1,n}$ are order statistics of the sample $\{Z_i\}$. Plotting $\hat{\xi}$ against k, we find that the value of $\hat{\xi}\approx0.6$ (95% CI: [0.4,1.0]) observed for a broad range of k is similar to the estimates above (Fig.1E of the main text). We found similar values of $\hat{\xi}$ for two other estimators, the Pickands and Dekkers-Einmahl-de Haan estimators (1,2).

Finally, we note here that a negative binomial distribution of Z, with its exponential tail, would have predicted the distribution of SSEs to be Gumbel-like if each SSE were indeed a maximum of samples of Z. This assertion can be proven by verifying the conditions

$$\lim_{n\to\infty}\frac{\sum_{n}^{\infty}P_j}{\sum_{n+1}^{\infty}P_j}=\text{const.},\qquad\lim_{n\to\infty}\sum_{n+2}^{\infty}\frac{P_j}{P_{n+1}}-\sum_{n+1}^{\infty}\frac{P_j}{P_n}=0,$$

where $P_j=\Pr(Z=j)$, sufficient for any discrete distribution to lie in a Gumbel-like domain of attraction (3). Thus, these considerations provide additional evidence suggesting that Z is not negative binomial.

SI References

1. Embrechts, P., Klüppelberg, C., and Mikosch, K. *Modelling Extremal Events for Insurance and Finance.* Springer Stochastic Modelling and Applied Probability (1997).
2. Wong, F. *et al.,* Supporting code for the paper available online at https://github.com/felixjwong/superspreaders.
3. Anderson, C. W. Local limit theorems for the maxima of discrete random variables. *Math. Proc. Camb. Phil. Soc.* **88**, 161-165 (1980).