Supporting Information Appendix: Echo chambers in the age of misinformation

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni,
Antonio Scala, Guido Caldarelli, H. Eugene Stanley, Walter Quattrociocchi

1 Data Description

Using the approach described in Ref. [1], we define the space of our investigation with the support of diverse Facebook groups that are active in the debunking of conspiracy theories.

The resulting dataset is composed of 67 public pages divided between conspiracy theories and science news. A second set, composed of two troll pages, is used as a benchmark to fit our data-driven model. The first category (conspiracy theories) includes the pages that disseminate alternative, controversial information, often lacking supporting evidence and frequently advancing conspiracy theories. The second category (science news) includes the pages that disseminate scientific information. The third category (trolls) includes those pages that intentionally disseminate false information on the Web.

For the three sets of pages we download all the posts (and their respective user interactions) across a five-year timespan (2010 to 2014). We perform the data collection process by using the Facebook Graph API [2], which is publicly available and accessible through any personal Facebook user account. The exact breakdown of the data is presented in Table 1.

2 Statistical Tools

Wald Test. We use the Wald test to compare the scaling parameters of two power law distributions. We define it as

$$H_0 : \hat{\alpha_1} = \hat{\alpha_2}$$
$$H_1 : \hat{\alpha_1} \neq \hat{\alpha_2}$$

	Total	Science	Conspiracy	Troll
Pages	69	35	32	2
News	9,642	5,032	3,538	1,072
Labeled Users	73,379	14,613	58,766	–
Shares	266,211	59,059	181,914	25,238

Table 1: Data Description.

where $\hat{\alpha_1}$ and $\hat{\alpha_2}$ are the estimated scaling parameters. The Wald statistic:

$$W = \frac{(\hat{\alpha_1} - \hat{\alpha_2})^2}{Var(\hat{\alpha_1})},$$

follows a χ^2 distribution with one degree of freedom. We reject the null hypothesis H_0 and conclude that there is a significant difference between the two scaling parameters if the p-value of W is below a given significance level α.

Kolmogorov-Smirnov Test. We use the Kolmogorov-Smirnov test to compare the empirical distribution functions of two samples. The Kolmogorov-Smirnov statistic for two given cumulative distribution functions $F_1(x)$ and $F_2(x)$ is

$$D = \sup_x |F_1(x) - F_2(x)|,$$

which measures the maximum punctual distance between the two sample distributions. If D is bigger than a given critical value D_α[1] we reject the null hypothesis $H_0 : F_1(x) = F_2(x)$ and conclude that there is a significant difference between the two sample distributions.

3 Anatomy of Cascades

3.1 Basic Properties

We studied the basic properties (size, height, max degree, and mean degree) of sharing trees for the three categories –i.e., science news, conspiracy theories, and trolling. Figure 1 summarizes these properties showing the CCDF of size (Fig. 1(a)), the CDF of height (Fig. 1(b)), the CCDF of maximum degree (Fig. 1(c)), and that of mean degree (Fig. 1(d)) for all categories. We note that both size and maximum degree are power low distributed and that the behavior of the different measures is similar for all categories. We estimated the power law exponents for both measures on the three sets, they are for size respectively 2.21, 2.47, 2.44 for science news, conspiracy theories, and trolling; while for max degree they are respectively 2.16, 2.45, 2.41. The maximum height reached is 5 for science news and conspiracy theories, and 4 for trolling, while for all categories there is a high probability that the height of the sharing tree remains below 3. The mean degree is with high probability much smaller than 10 for all categories.

Figure 2 shows the size as a function of lifetime[2] for science news (a) and conspiracy theories (b). We notice again, as in Figure 2 in the text, a contents-driven differentiation in the sharing patterns. The positive correlation between lifetime and size observed for conspiracy theories is confirmed, while for science news the dynamics is slightly more complex.

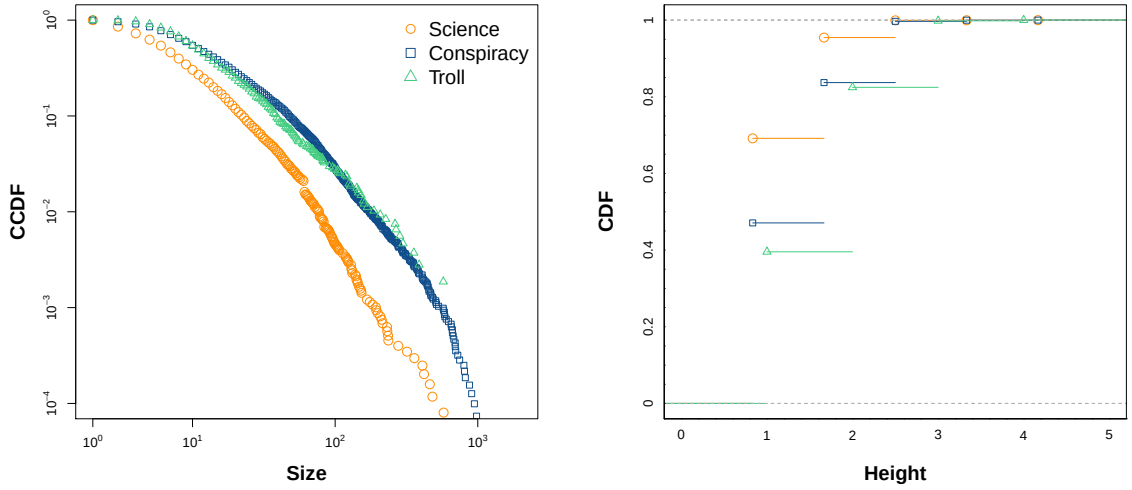
We computed the probability density function (PDF) of the edge homogeneity, Figure 3, together on science news and conspiracy theories (a), for the unconditional case and for the conditional one on the event that the user that made the first share in the couple has a positive or negative

[1] The critical value D_α depends on the sample sizes and on the considered significance level α, it can be computed as

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

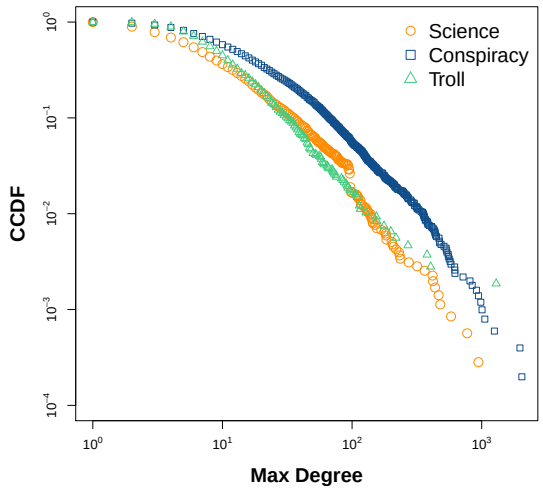
where n_1 and n_2 are the respective sample sizes and $c(\alpha)$ is a fixed value associated with the significance level α.

[2] We recall that the lifetime of each post is here defined as the temporal distance, in hours, between the first and last share of the post.

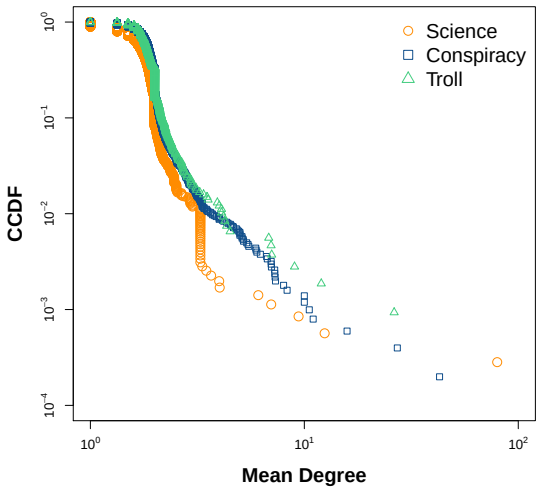


(a)

(b)



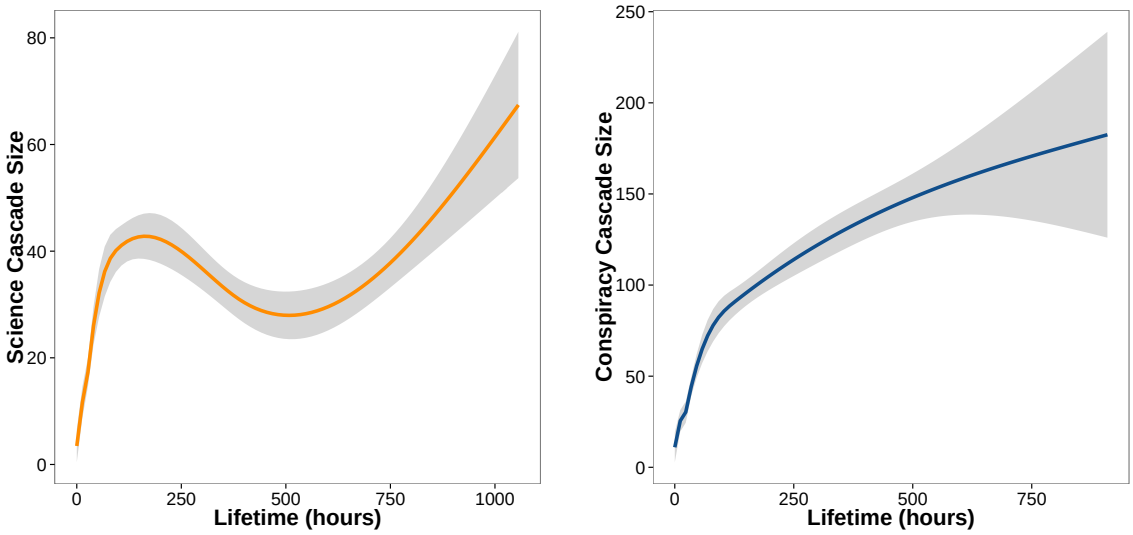
(c)



(d)

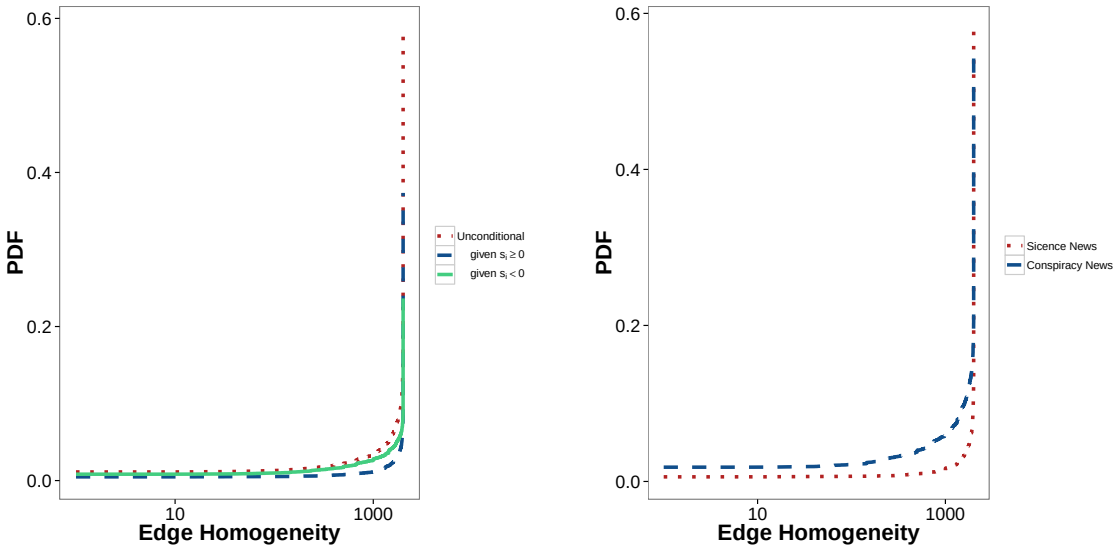
Figure 1: CCDF of Size (a), CDF of Height (b), CCDF of Maximum Degree (c), and CCDF of Mean Degree (d). Size and max degree show power law distributions, where the estimated exponents for the power law distribution of size are 2.21, 2.47, 2.44 and those of max degree are 2.16, 2.45, 2.41, respectively for science news, conspiracy theories, and trolling. Height is generally low, with the maximum level being 5 for science news and conspiracy theories, and 4 for trolling.

polarization, and for science and conspiracy separately (b), unconditional case only. Note that both on science and conspiracy the probability of a negative mean edge homogeneity is zero, indicating a strong dominance of homogeneous links with respect to non homogeneous ones.



(a) (b)

Figure 2: Cascade size as a function of lifetime, for science news (a) and conspiracy theories (b). We notice a contents-driven differentiation in the sharing patterns.



(a) (b)

Figure 3: Probability density function (PDF) of edge homogeneity. In (a) the PDF is computed together on science news and conspiracy theories for the unconditional case (dotted red) and for the conditional case on the event that the user that made the first share in the couple has a positive (dashed blue) or negative (solid green) polarization. In (b) only the unconditional case is reported, computed separately on science news (dotted red) and conspiracy theories (dashed blue).

3.2 Homogeneity Evidences

It is often debated if social activities and viral phenomena are more influenced by the social structure or by the similarity of the individuals. We investigated this aspect by considering the underlying

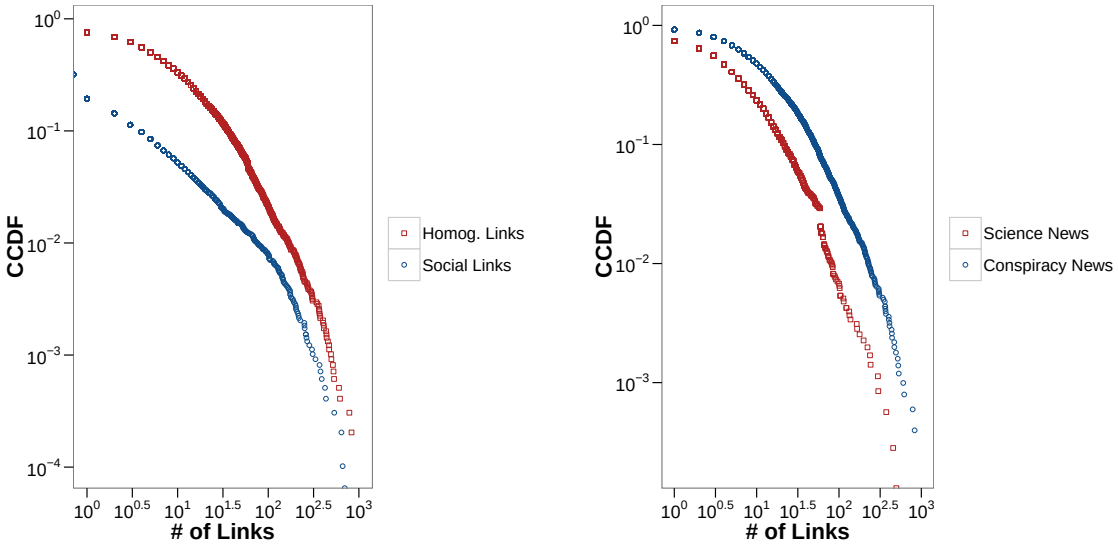
social structure of our network: we refer to the official friendship links that users share on Facebook as social links; while we say that two successive nodes in the sharing tree have a homogeneous link if their user polarization signs are concordant. More formally, an edge of the sharing tree is said to be a homogeneous link if its edge homogeneity is positive. Notice that homogeneous links only occur between nodes linked in the sharing tree, while social ones may occur between any two nodes in the sharing tree.

We computed the complementary cumulative distribution function (CCDF) of the number of social and homogeneous links in each sharing tree, with no category distinction (Figure 4 (a)), and that of number of homogeneous links separately on science news and conspiracy theories (Figure 4 (b)). We compared the two couple of distributions by the Wald test, with the null hypothesis H_0 that the two scaling parameters are equal, and with significance level $\alpha = 0.05$. Estimated parameters and p-values are reported in Table 2. In both cases the p-value is much smaller than

	$\hat{\alpha_1}$	$\hat{\alpha_2}$	p
Case I	2.606	1.853	0
Case II	2.371	2.542	5×10^{-12}

Table 2: Results from Wald test, where α_1 and α_2 are the two estimated scaling parameters for each couple and p is the corresponding p-value.

the significance level and the two distribution parameters are significantly statistically different, we then reject the null hypothesis. We note that the distribution of homogeneous links has a similar behavior for both categories and that the probability to find homogeneous links is generally higher than that of social links, indicating a prominent role of homogeneity with respect to the network structure.



(a) (b)

Figure 4: Complementary cumulative distribution function (CCDF) of the number of social and homogeneous links on the two sets of news together (a), and of the number of homogeneous links on science and conspiracy news separately (b).

Figure 5 shows the complementary cumulative distribution function (CCDF) of the number

of homogeneous and total paths, for the three following samples: science news and conspiracy theories together (left), science news (center), and conspiracy theories (right). More formally, we considered, for each tree, the number of all sharing paths from the root to one of the leaves, and we compared it with the number of sharing paths with positive edge homogeneity, meaning that both edge's endpoints show a user polarization of the same sign, we call them homogeneous paths.

Looking at Figure 5 we notice a high similarity for all the couples, for this reason we compare them by using Wald test and Kolmogorov-Smirnov test, with level of significance $\alpha = 0.01$. The null hypothesis is the equivalence of the two scaling parameters for the Wald test and the equivalence of the whole sample distributions for the Kolmogorov-Smirnov. Table 3 reports the results from Wald test, while Table 4 those from Kolmogorov-Smirnov test.

We fail to reject the null hypothesis of Wald test in the second and third case, i.e., science news and conspiracy theories separately, while we reject it in the case of the whole sample. However, we fail to reject the null hypothesis of Kolmogorov-Smirnov in all three cases, as the maximum distance is always smaller than the critical value, we deduce that the distributions are not significantly statistically different in all three cases, and the same is true for the scaling parameters in the case of science news and conspiracy theories separately.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	p
Case I	2.037	2.089	$8.45 \,\mathrm{x}\, 10^{-7}$
Case II	2.427	2.447	0.413
Case III	2.054	2.026	0.040

Table 3: Results from Wald test, where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the two estimated scaling parameters for each couple and p is the corresponding p-value.

	D	D_α	p
Case I	0.0216	0.0233	0.02047
Case II	0.0199	0.0378	0.4525
Case III	0.0262	0.0296	0.03204

Table 4: Results from Kolmogorov-Smirnov test. D is the estimated maximum distance between the two distributions under analysis, D_α is the corresponding critical value, and p the corresponding p-value.

To better visualize the similarity between each pair of distributions we show the Q-Q plots in Figure 6.

Figure 7 shows the frequency of maximum length for all sharing paths[3] and homogeneous paths for science news (left) and conspiracy theories (right). We confirm the pervasiveness of homogeneous paths, but we also find homogeneous paths in which there is a shift of -1 in the path length (with respect to the total path length k). These $(k-1)$-homogeneous paths may be caused by a discordant sharing in the first step (i.e., when the product of the first sharer's polarization and the sharer page category is negative), because all the following shares appear to be driven by homogeneity.

[3] A sharing path is here defined as any path from the root to one of the leaves of the sharing tree. A homogeneous path is a sharing path for which the edge homogeneity of each edge is positive

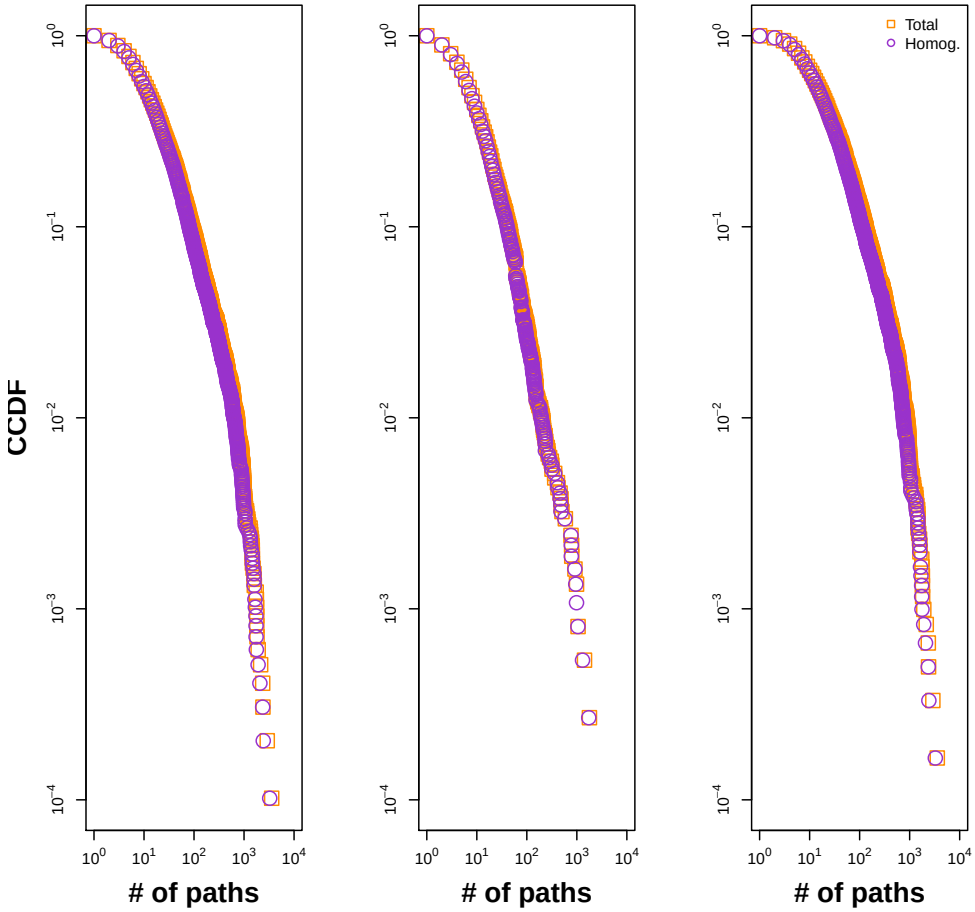
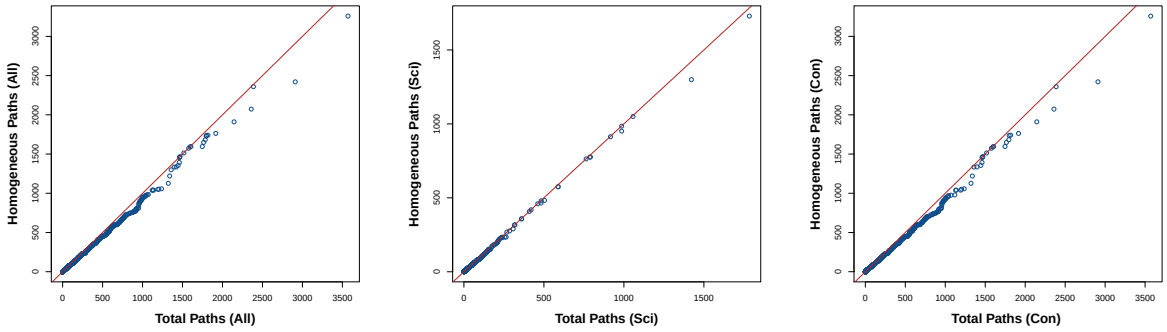


Figure 5: Complementary cumulative distribution function (CCDF) of the number of total and homogeneous paths on the whole science and conspiracy sample (left), on science news (center), and on conspiracy theories (right). Results from the Kolmogorov-Smirnov test on the three distributions couples, with the null hypothesis H_0 that the distributions in each couple are equal, show the following p-values: 0.02, 0.45, 0.03, leading us to reject the null hypothesis in all cases ($\alpha = 0.01$). On the other hand the maximum estimated distances are respectively 0.0216, 0.0199, 0.0262 and corresponding critical values 0.0232, 0.0377, 0.0296; the maximum distance is smaller than the critical value in all cases, so we fail to reject the null hypothesis. We can consider the distributions in each couple as equal, meaning that homogeneous paths are pervasive.

4 Model

4.1 First Sharers

A fundamental step in our data-driven percolation model is the first temporal share, hence we simulate the model's dynamics considering different fits of the data distribution of first shares on the whole sample of science news and conspiracy theories. Figure 8 shows the fit for different distributions: Inverse Gaussian (blue), Log Normal (violet), Poisson (orange), and Uniform in $(1, 100)$ (green). We notice that the best fit is obtained by the Inverse Gaussian with mean 39.34



(a) (b) (c)

Figure 6: Q-Q plots for the three couples of distributions in Figure 5: number of total vs homogeneous links on science news and conspiracy theories together (a), on science news (b), and on conspiracy theories (c). In all three cases the distributions can be considered as equal.

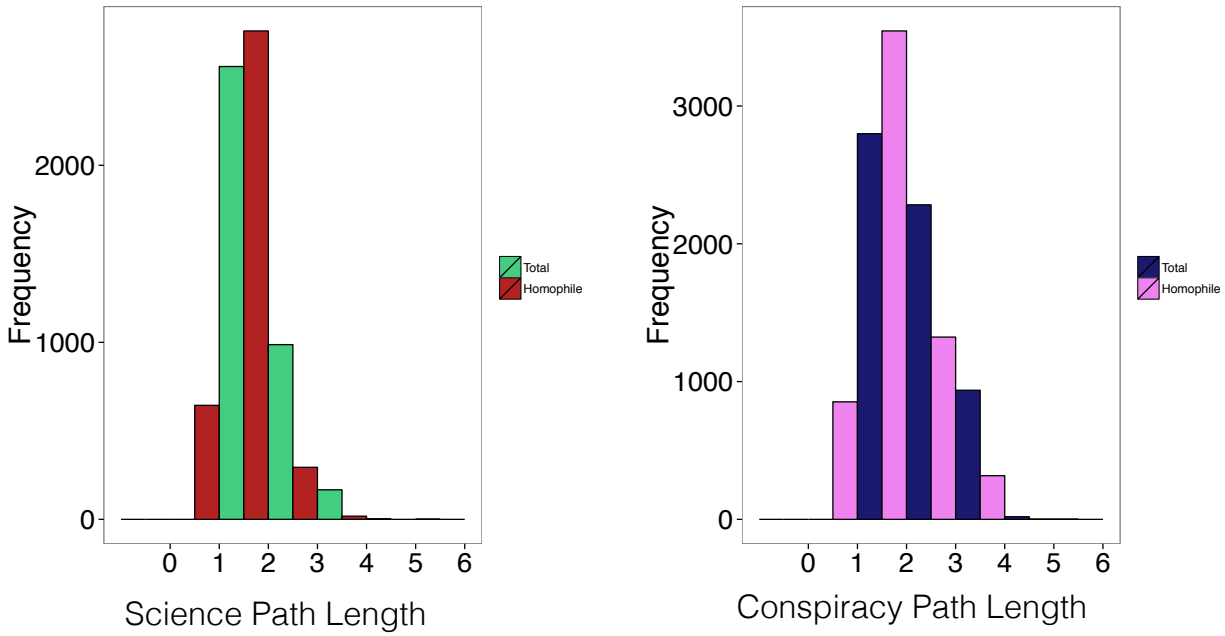


Figure 7: Frequency of maximum length for all paths and homogeneous paths, on science news (left) and conspiracy theories (right). The pervasiveness of homogeneous paths observed in Figure 5-6 is confirmed.

and scale parameter 6.28. Table 5 shows the estimated parameters of the fit of Inverse Gaussian, Log Normal, and Poisson both on the whole sample and on the troll news sample, these parameters are used in the simulations of the model and in the fit of the model on trolling messages, together with the baseline Uniform distribution.

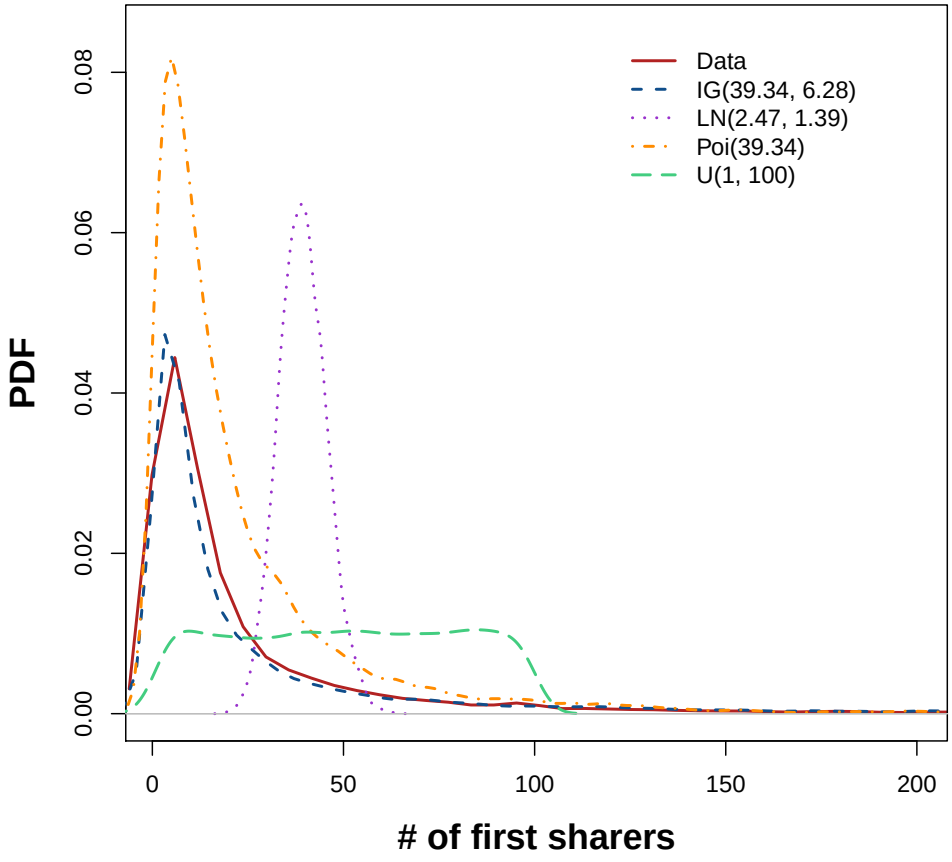


Figure 8: Fit of the probability distribution function of first sharers with different distributions: Inverse Gaussian (blue), Log Normal (violet), Poisson (orange), and Uniform in $(1, 100)$ (green).

4.2 Simulation Results

Figure 9 shows the results obtained simulating the model for $n = 5,000$ users, $m = 1,000$ news, and varying the fraction of homogeneous links $\phi_{HL} \in [0.5, 1]$ and the rewiring probability $r \in [0.01, 0.2]$; while Figure 10 shows the results for the fit of the model on the trolling set, where the number of users and the number of messages is taken from the trolling set data ($n = 16,889$ is the number of users active in the trolling category and $m = 1,072$ is the number of trolling messages in the dataset), and the parameters ϕ_{HL} and r vary in the same intervals as before.

Table 6 shows the combination of parameters that best reproduces the data, and Fig. 11 shows the average size (left) and average height (right) produced by the simulations, where different colors indicate the different distributions of the number of first sharers considered, and we vary the sharing threshold δ and the fraction of homogeneous links ϕ_{HL}, i.e., $\delta \in [0.01, 0.05]$ and $\phi_{HL} \in [0.5, 0.59]$.

Table 7 shows a summary of relevant statistics (min value, first quantile, median, mean, third quantile, and max value) to compare the real data first sharers distribution with the fitted distributions. The Inverse Gaussian (IG), with mean 18.73 and scale parameter 9.63, shows the best fit for the distribution of first sharers from trolling category, with respect to all the considered statistics.

	Science and Conspiracy	Troll
IG	$(39.34, 6.28)$	$(18.73, 9.63)$
LN	$(2.47, 1.39)$	$(2.21, 0.93)$
Poi	39.3	18.73

Table 5: Estimated parameters for the Inverse Gaussian, Log Normal, and Poisson distribution obtained fitting each distribution with the distribution of first sharers of science and conspiracy categories in the same sample and of trolling category on its own. Both samples are taken from the data.

Parameters.		Size		Height	
	(ϕ_{HL}, δ)	mean	std deviation	mean	std deviation
Real Data	-	23.54	122.32	1.78	0.73
Real Data Simulated	.56, .015	23.52	133.02	1.28	1.18
IG Simulated	.56, .015	23.42	33.43	1.28	0.88

Table 6: Mean cascades size and height obtained with the best parameter combination compared to real data measures. Simulation results are reported for two cases: number of first sharers distributed as real data and as inverse Gaussian.

Taking into account the result in Table 7, we simulated the model dynamics with the first sharers distributed as $IG(18.73, 9.63)$. CDF of height and the CCDF of size for the best combination of model parameters $(\phi_{HL}, r, \delta) = (0.56, 0.01, 0.015)$ is reported in Figure 5 in the main text. A Kolmogorov-Smirnov test on the two size distributions yields a maximum estimated distance $D = 0.1371$ (critical value 0.059, $\alpha = 0.01$) and a p-value $p = 3.52 \times 10^{-9}$, hence we reject the null hypothesis. Table 8 shows a summary of relevant statistics (min value, first quantile, median, mean, third quantile, and max value) to compare the real data size and height distributions with the fitted ones, for the same distribution of first sharers and the same parameters combination. We notice that the fit is good for all the statistics, with the exception of min and max value of size. For the min value, the presence of a zero is due to the fact that the inverse Gaussian is a real valued distribution function and in the simulations we considered the integer part of the number of first sharers, thus producing a number of never shared pieces of information. On the other hand, the high difference in the max value is probably due to the long tail of the data size distribution.

	Data	IG	LN	Poi
Min	*1*	**0.81**	0.16	7
1st Qu.	*5*	**4.67**	2.13	16
Median	*8*	**9.71**	4.66	19
Mean	*18.73*	**18.73**	9.99	18.72
3rd Qu.	*16*	**21.76**	11.85	22
Max	*3882*	**346.60**	183.40	32

Table 7: Summary of relevant statistics for the first sharers distribution fitted with an Inverse Gaussian, a Lognormal, and a Poisson distribution.

		Size		Height
	Data	Simulated	Data	Simulated
Min	2	0	1	1
1st Qu.	7	6.19	1	2
Median	10	11.93	2	2
Mean	23.54	23.84	1.78	2.18
3rd Qu.	19.25	27.17	2	3
Max	3845	541.80	4	5

Table 8: Summary of relevant statistics for size and height distributions.

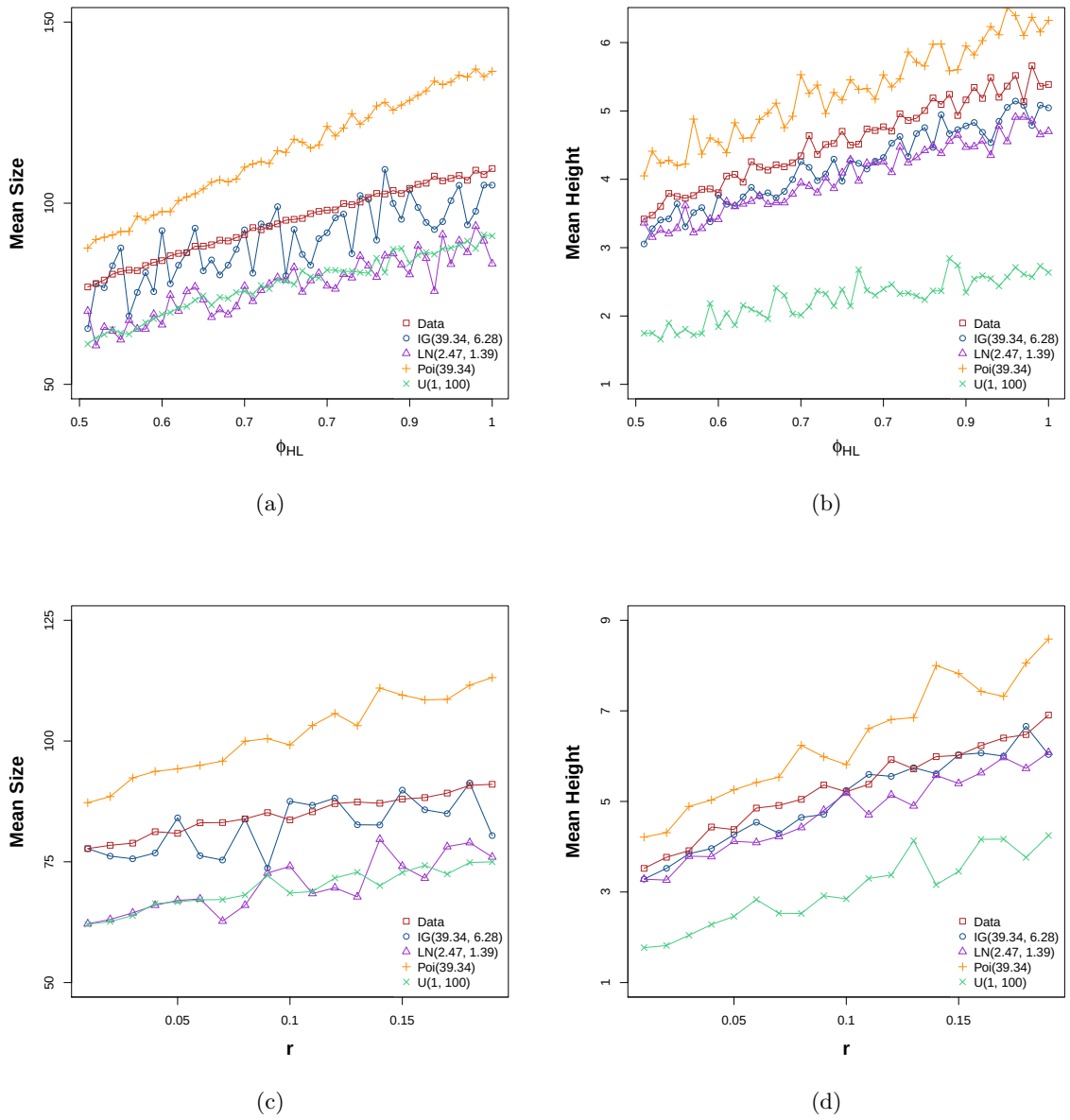


Figure 9: Simulation results for the mean size (a,c) and the mean height (b,d) with $n = 5,000$ users, $m = 1,000$ news, fixed sharing threshold $\delta = 0.05$ and different combinations of fraction of homogeneous links ϕ_{HL} and rewiring probability r. In part (a) and (b) $r = 0.01$ is fixed and ϕ_{HL} varies in the interval $[0.5, 1]$; while in part (c) and (d) $\phi_{HL} = 0.5$ is fixed and r varies in the interval $[0.01, 0.2]$. The different colors and shapes of the curves indicate the different distributions of first sharers used in the simulations: red square is for the sample data distribution, blue circle for the Inverse Gaussian $IG(39.33, 6.27)$, violet triangle for the Lognormal $LN(2.46892, 1.39399)$, orange cross for the Poisson $Poi(39.3385)$, and green x for the Uniform $U(1, 100)$.

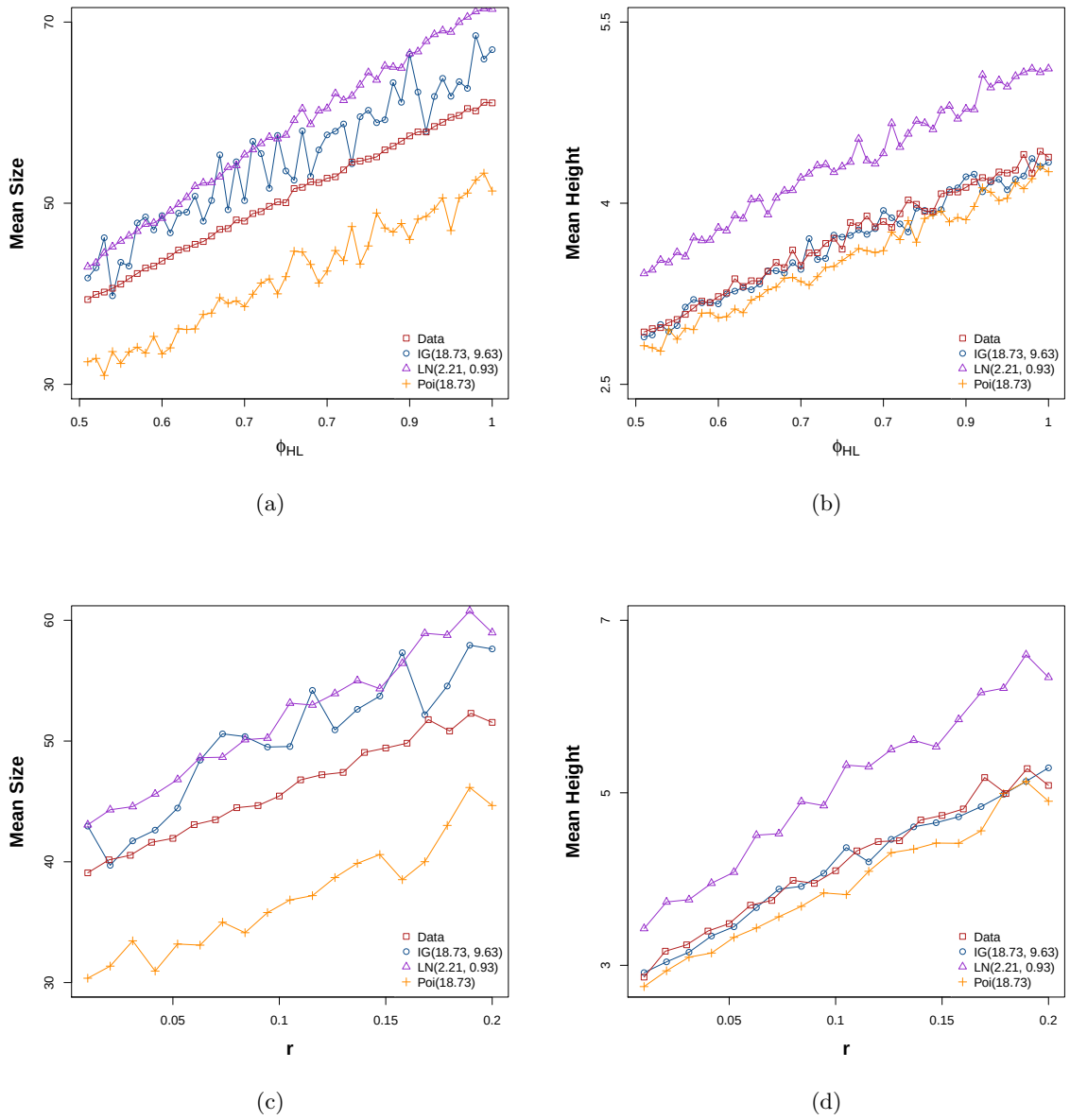


Figure 10: Fit of the model with trolling messages data, simulation results for the mean size (a,c) and the mean height (b,d) with $n = 16,889$ users, $m = 1,072$ news, fixed sharing threshold $\delta = 0.05$ and different combinations of fraction of homogeneous links ϕ_{HL} and rewiring probability r. In part (a) and (b) $r = 0.01$ is fixed and ϕ_{HL} varies in the interval $[0.5, 1]$; while in part (c) and (d) $\phi_{HL} = 0.5$ is fixed and r varies in the interval $[0.01, 0.2]$. The different colors and shapes of the curves indicate the different distributions of first sharers used in the simulations: red square is for the sample data distribution, blue circle for the Inverse Gaussian $IG(18.73, 9.63)$, violet triangle for the Lognormal $LN(2.21, 0.93)$, and orange cross for the Poisson $Poi(18.73)$.

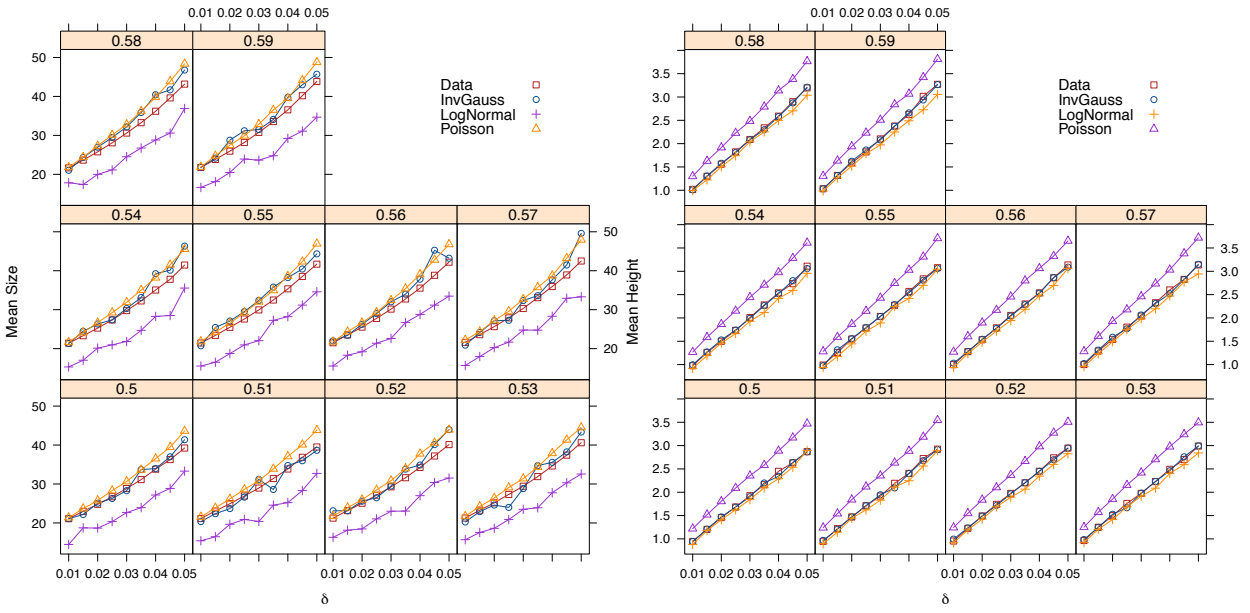


Figure 11: Fit of the model with trolling messages data, simulation results for the mean size (left) and the mean height (right) with $n = 16,889$ users, $m = 1,072$ news, fixed rewiring probability $r = 0.01$, fraction of homogeneous links ϕ_{HL} varying in the interval $[0.5, 0.59]$, and sharing threshold δ varying in $[0.01, 0.05]$. The different colors and shapes of the curves indicate the different distributions of first sharers used in the simulations.

References

[1] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: collective narratives in the age of (mis)information. *Plos ONE*, 2015.

[2] Facebook. Using the graph api. Website, 8 2013. last checked: 19.01.2014.