The spread of information on social media

Lília Perfeito, Pedro Duarte, Joana Gonçalves-Sá

Social Physics and Complexity @ LIP

FARE
Fake News and Real People: Using big data to understand human behaviour

About me

2004-2008

PhD in Biology

The distribution of fitness effects of spontaneous mutations in bacteria

2008-2012

Collaboration between Institute for genetics and Institute for theoretical physics

Empirical molecular fitness landscapes

2012-2018

Principal Investigator

Can we predict evolution?

2019-2020

Data Science and Policy

Social Physics & Complexity

About us

HEALTH	Online vs. Offline Patterns Emergency Now-casting Antibiotic Over-prescription	Google Trends SNS24 Twitter ER acceptance /times SPMS e-prescriptions	**Math Modelling** **ML** **Epidemiology**
POLICY	Political Decisions Gender Differences Agenda Setting Voting vs. Discourse	Media records Twitter Facebook Parliament data	**NLP** **Networks** **Math Modelling** **Complex Systems**
BEHAVIOUR	Cognitive Biases Attitudes Towards Science Tracking Anxiety	Large scale surveys Behavioral experiments Twitter Facebook	**Networks** **Math Modelling** **Psychology** **Information**

https://socialcomplexity.eu/

The spread of information on social media: a model based on evolutionary principles

Lília Perfeito, Pedro Duarte, Joana Gonçalves-Sá

Social Physics and Complexity @ LIP

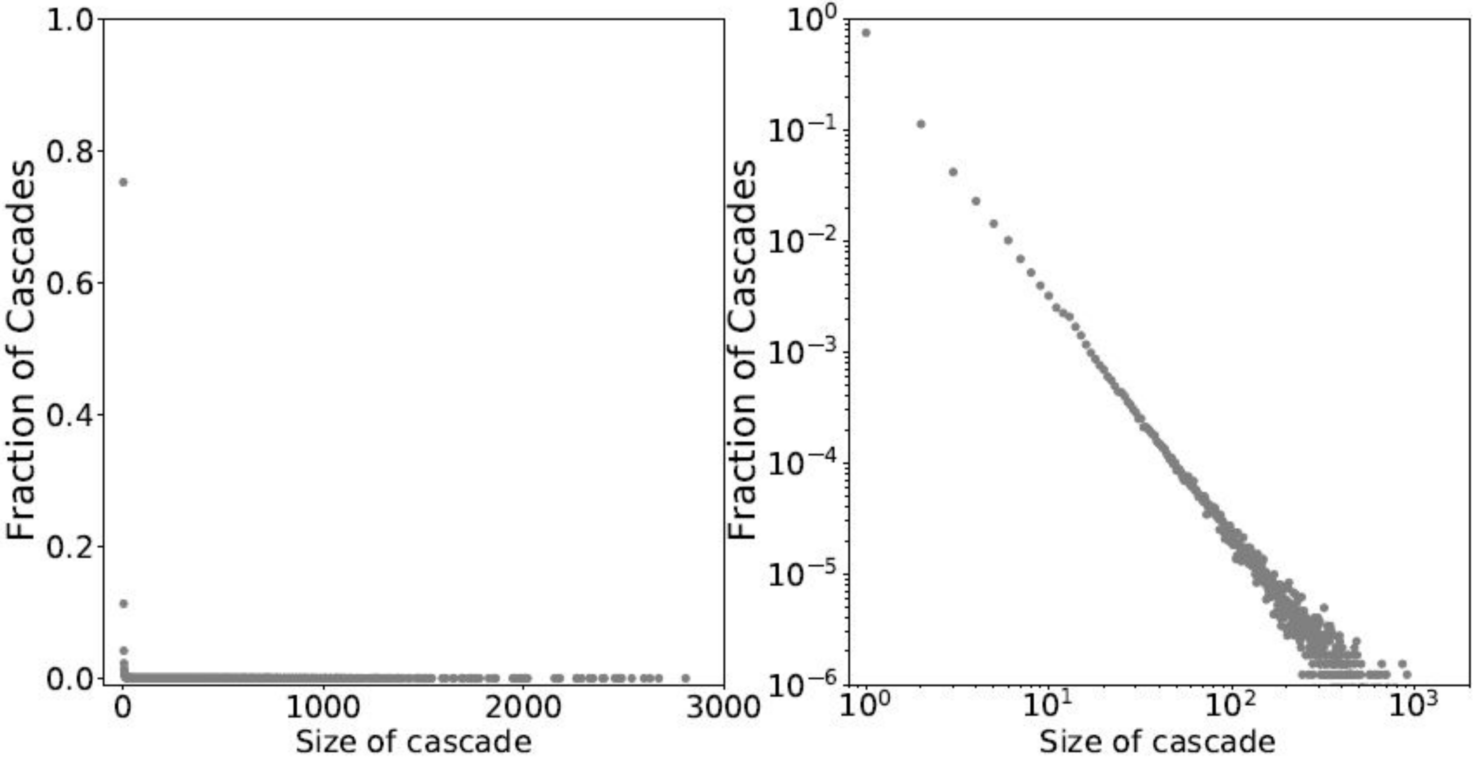
FARE
Fake News and Real People: Using big data to understand human behaviour

How does information spread on social networks?

Image credit: all-free-download.com

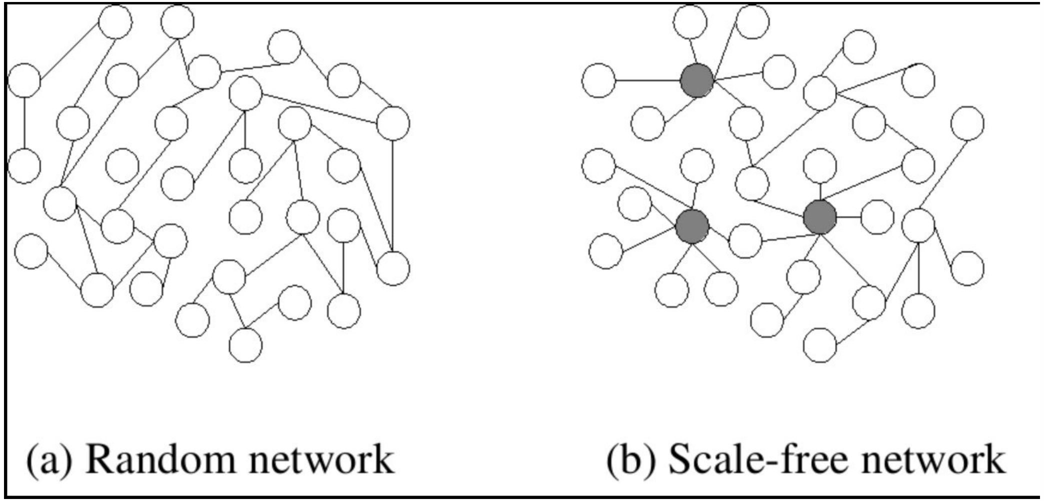
Lília Perfeito

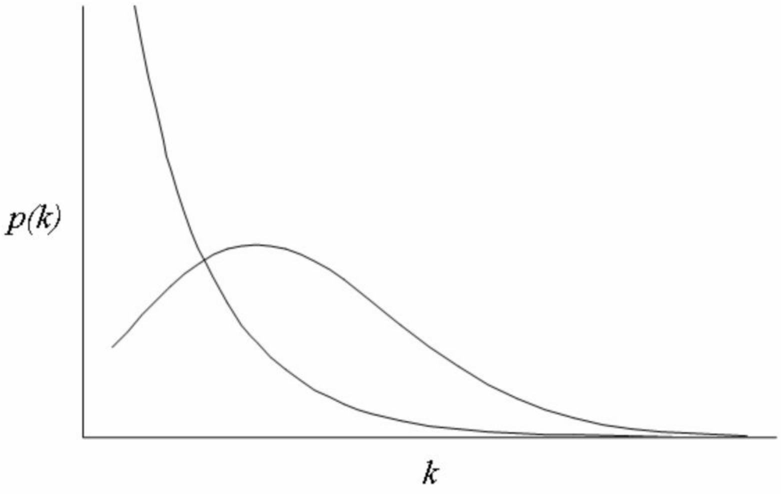
The success of information greatly varies

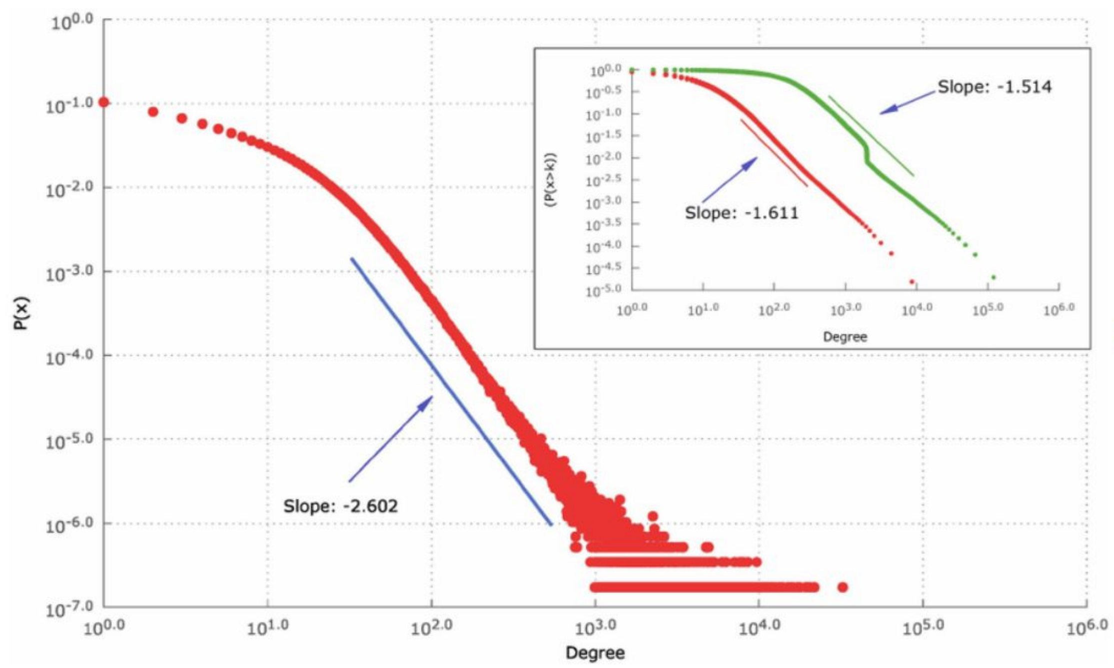


What are the main forces behind the success/failure of memes?

H1: The social network

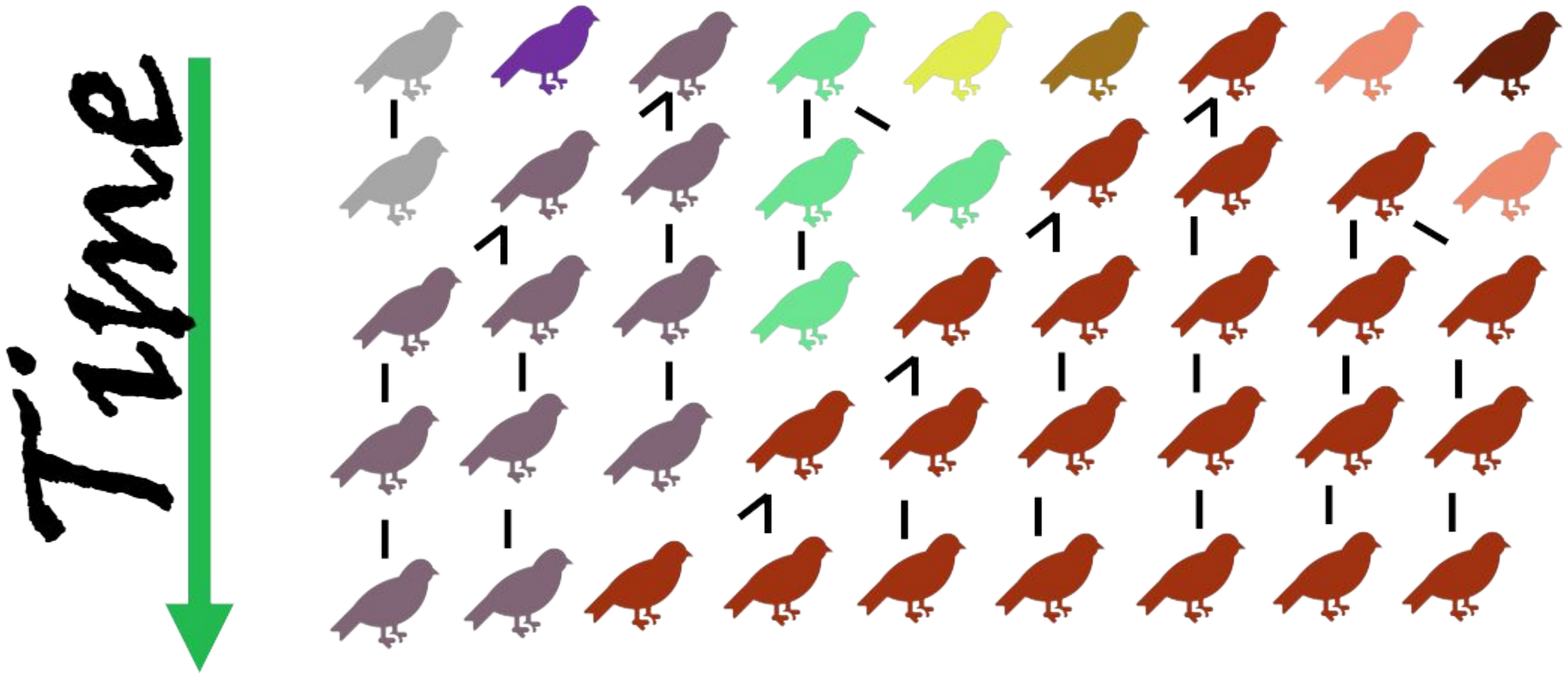




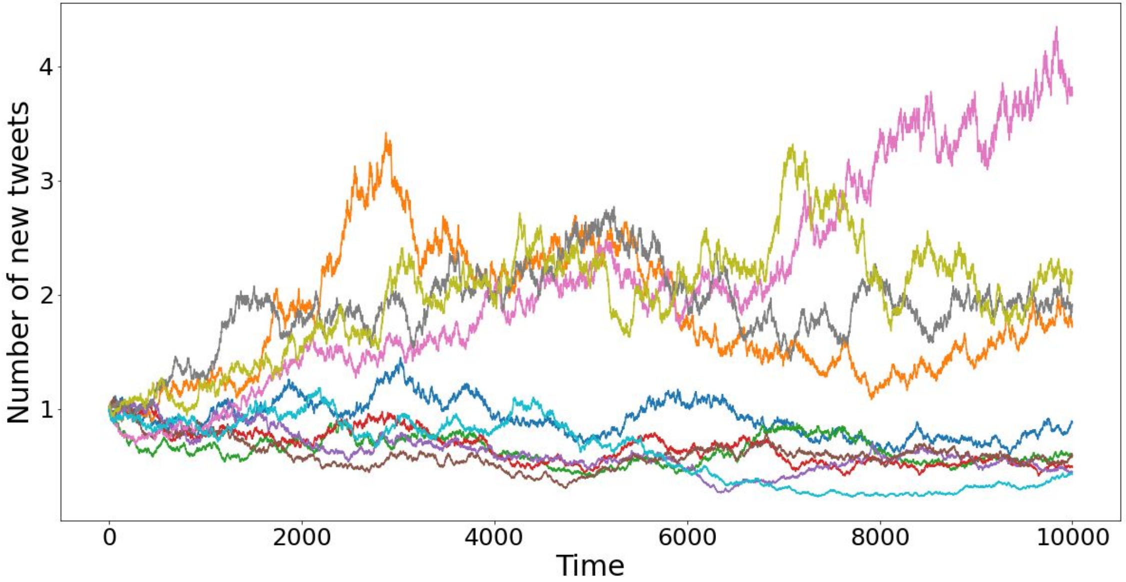


Szüle, J., Kondor, D., Dobos, L., Csabai, I., & Vattay, G. (2014). Lost in the city: revisiting Milgram's experiment in the age of social networks. *PloS one*, *9*(11), e111973.

H2: Exponential growth and selection



H2: Exponential growth and selection



$$\frac{dN}{dt} = x \cdot \epsilon$$

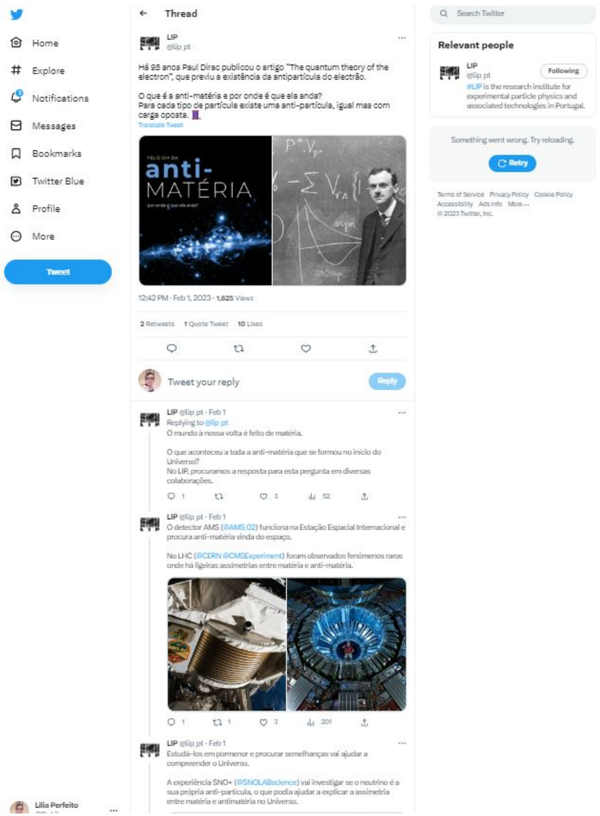
$$\frac{d\log(N)}{dt} = log(x) + \log(\epsilon)$$

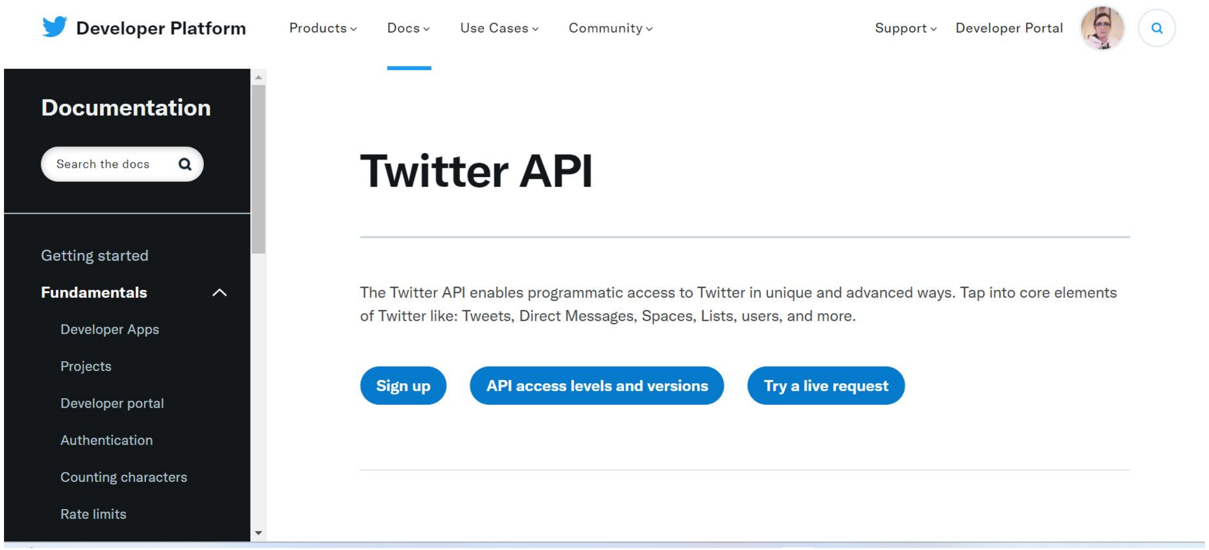
where log(ε) is a Gaussian random variable

Can we tease apart the role of the network and growth?

Can we measure selection acting on information?

Social media platforms - Twitter as a research tool





Lília Perfeito

The data

	Tweets + re-tweets		
Language	Covid	Music	Film
Portuguese	66 185 221	24 375 572	2 928 429
Italian	6 738 712	1 016 834	995 971
German	13 566 605	447 831	444 034
Dutch	7 159 327	191 366	186 972

The data

Language	Tweets + re-tweets		
	Covid	Music	Film
Portuguese	66 185 221	24 375 572	2 928 429
Italian	6 738 712	1 016 834	995 971
German	13 566 605	447 831	444 034
Dutch	7 159 327	191 366	186 972

April 2020 to June 2021

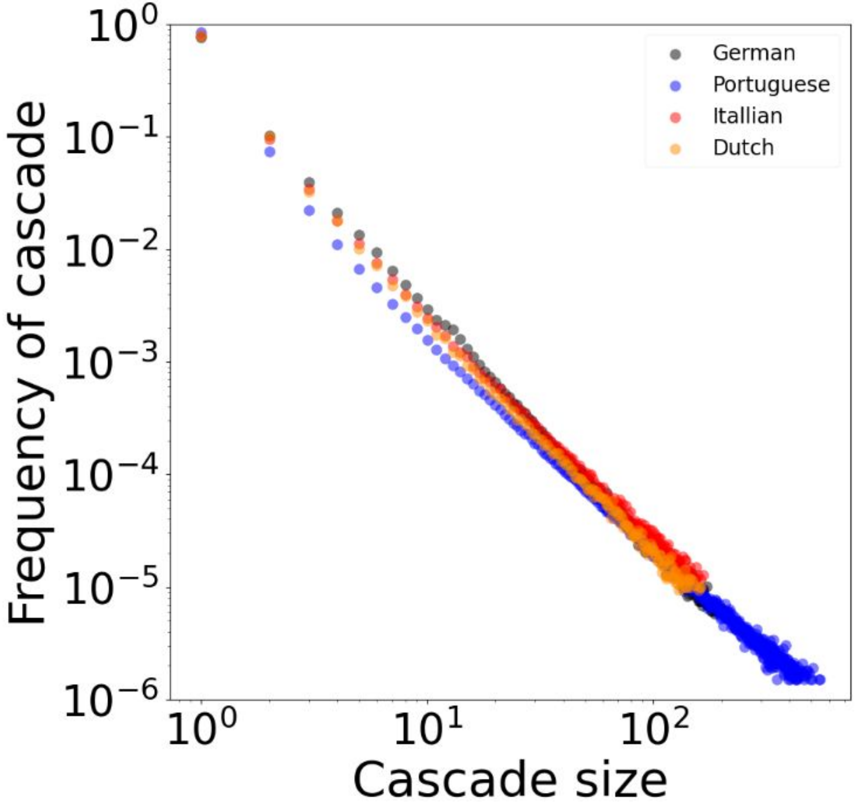
Language	Accounts			Original tweets		
	Covid	Music	Film	Covid	Music	Film
Portuguese	6 914 262	3 905 141	2 928 429	20 595 795	13 251 392	7 888 742
Italian	728 279	728 279	728 279	2 719 486	490 764	638 865
German	808 096	154 463	170 588	5 392 367	324 803	297 891
Dutch	461 353	69 203	95 396	3 293 620	138 851	115 969

Important definition

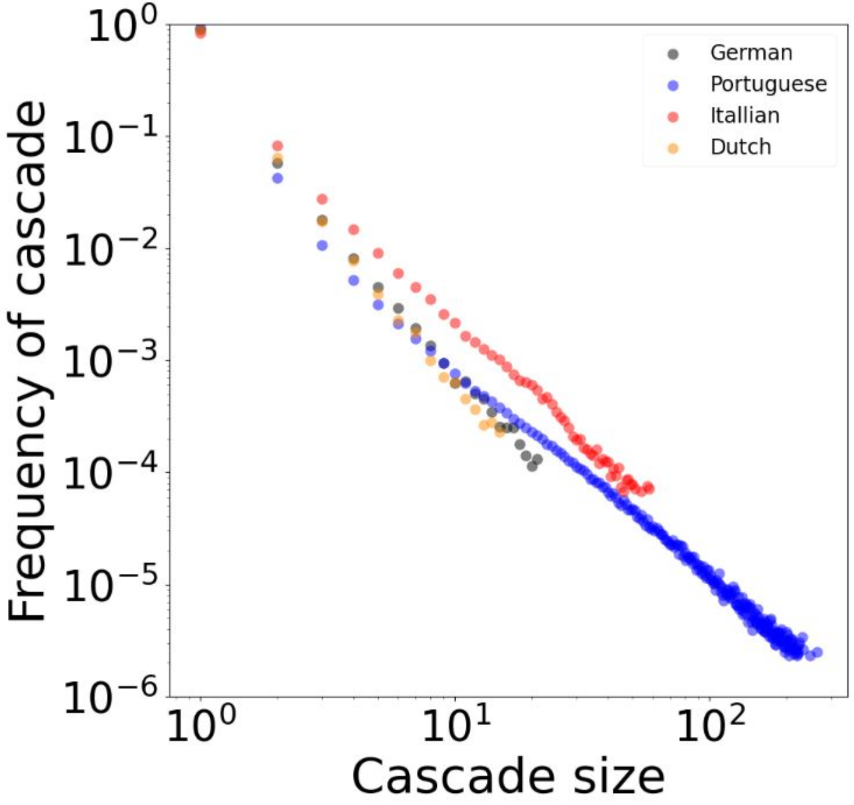
Cascade: Set of identical tweets, probably originating in a single seeding event.

The data

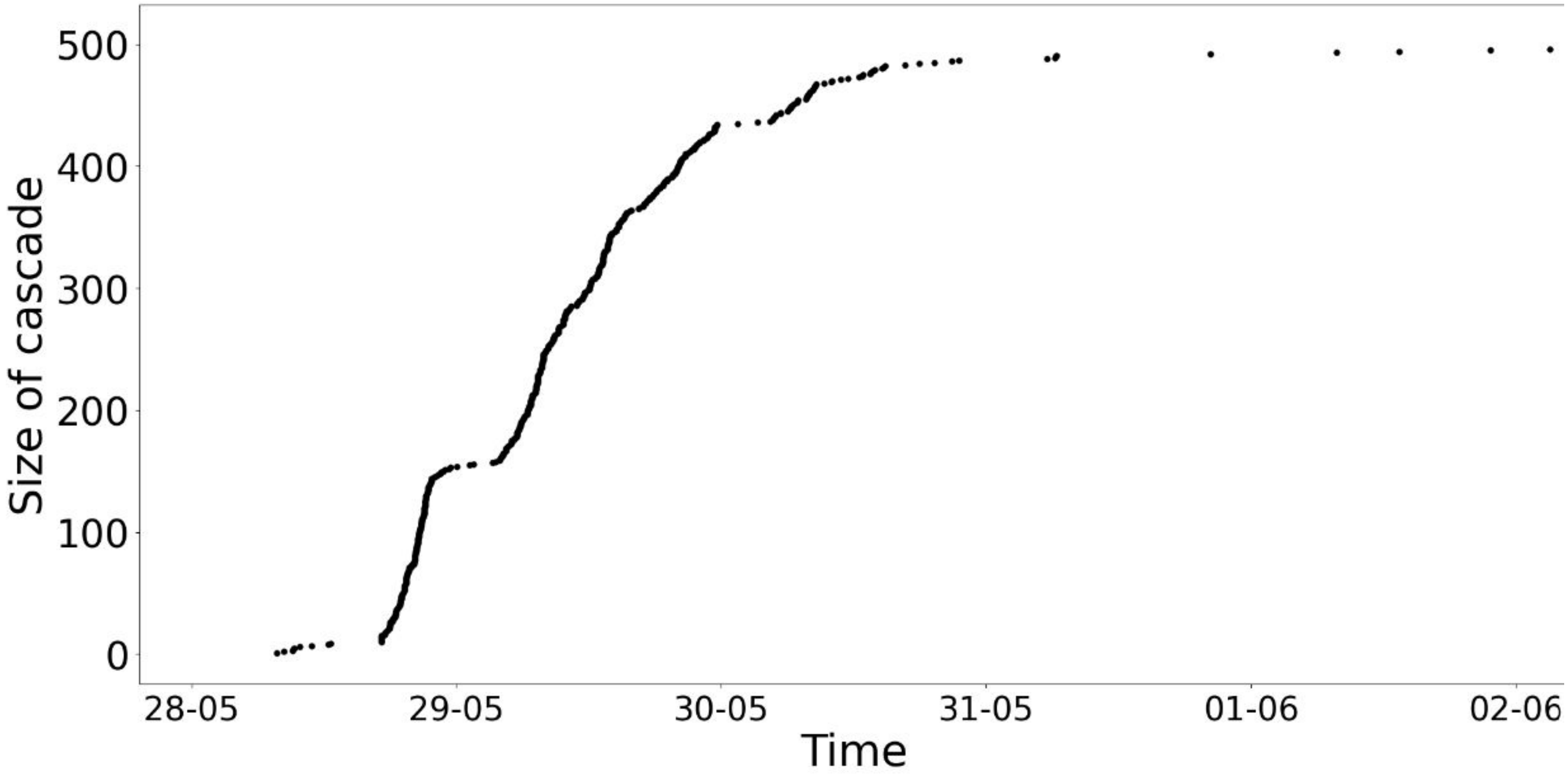
'Covid' or 'Corona'



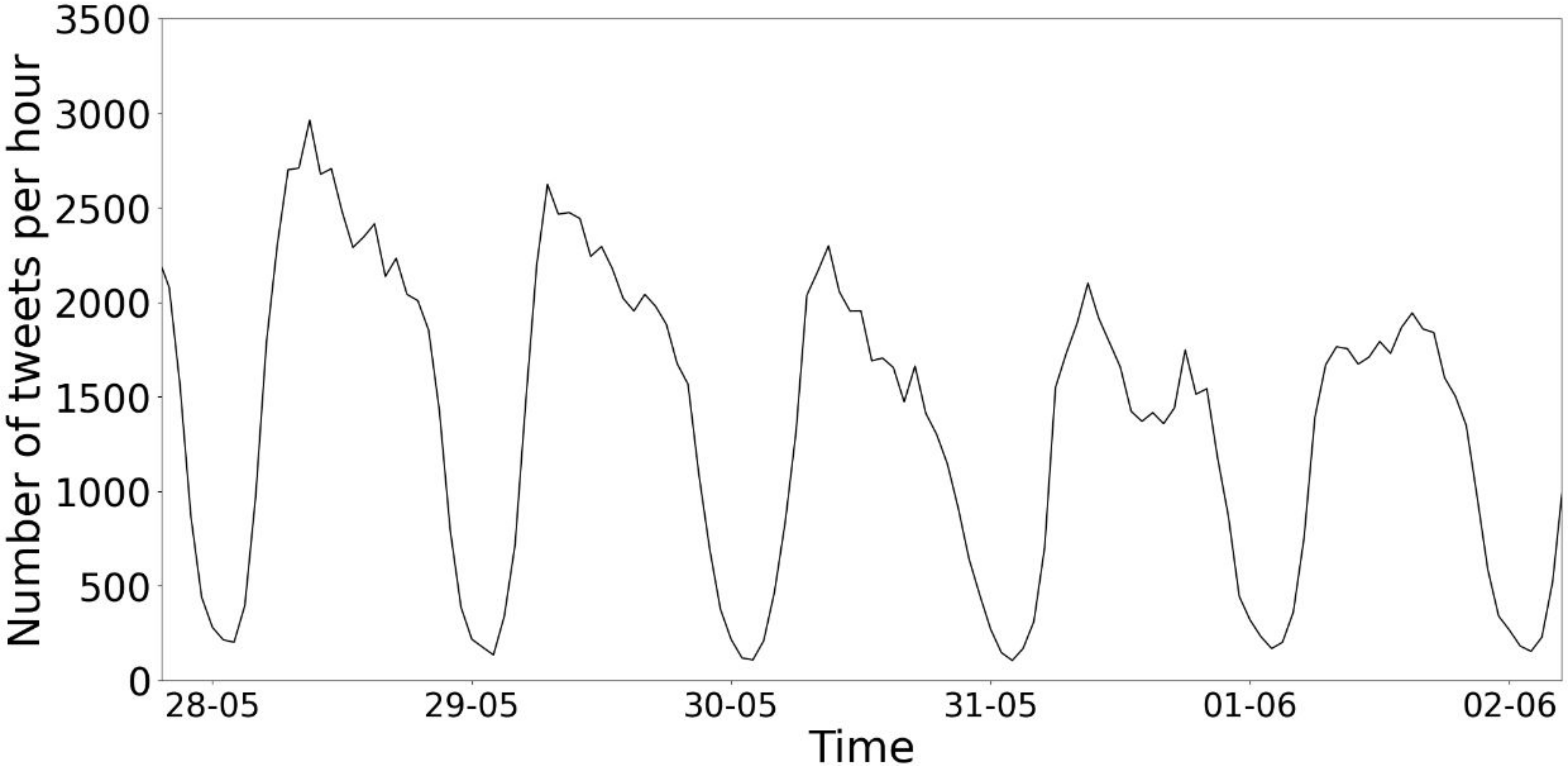
'Music'

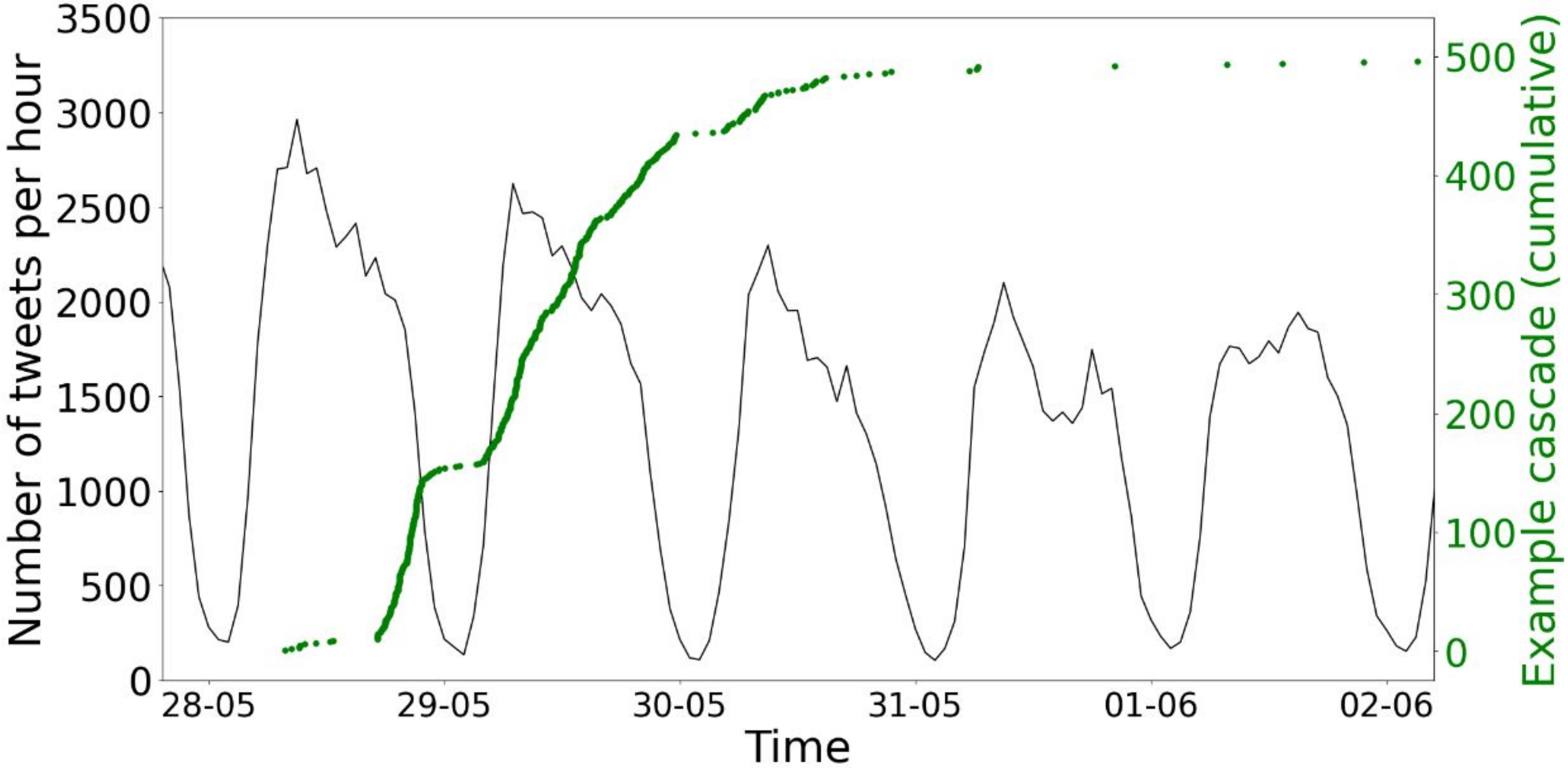


Twitter time

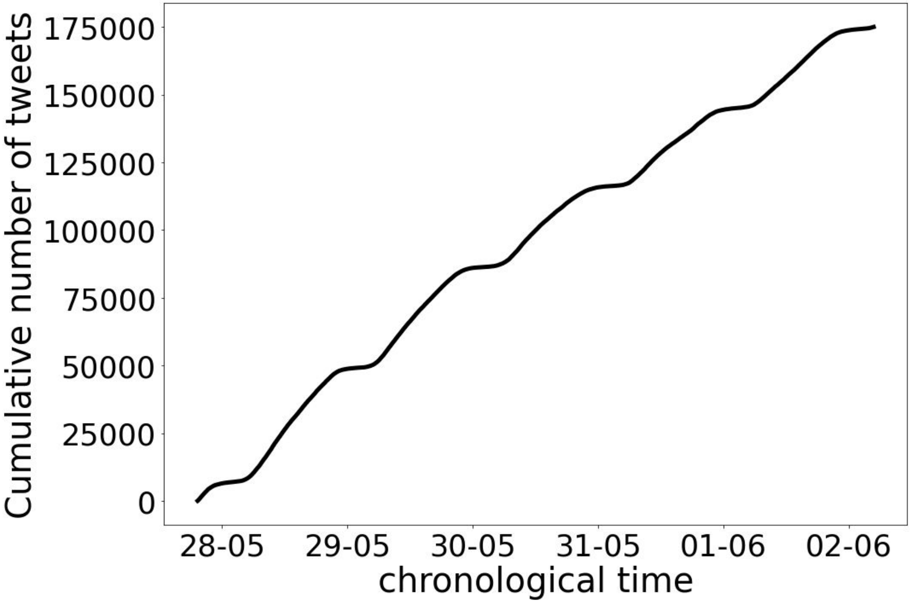


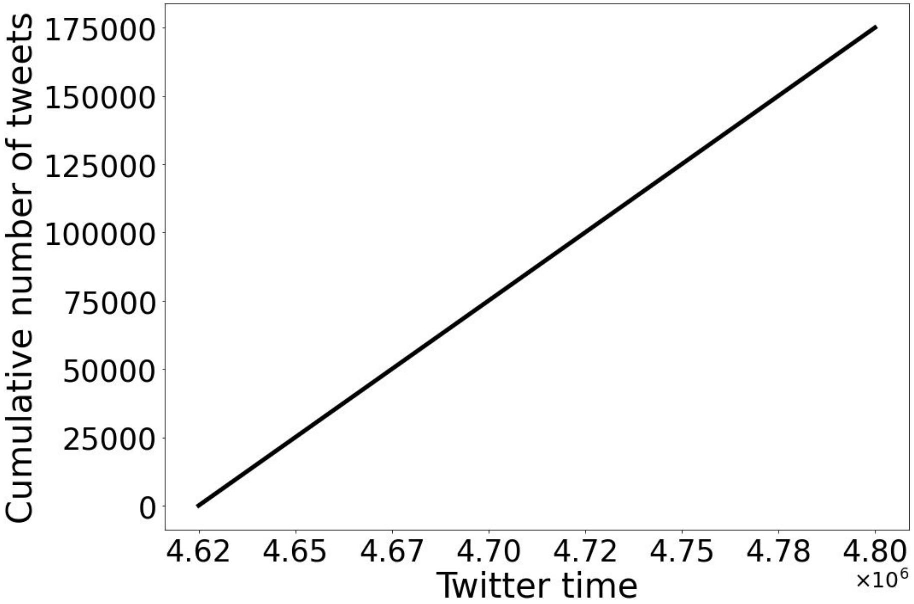
Twitter time



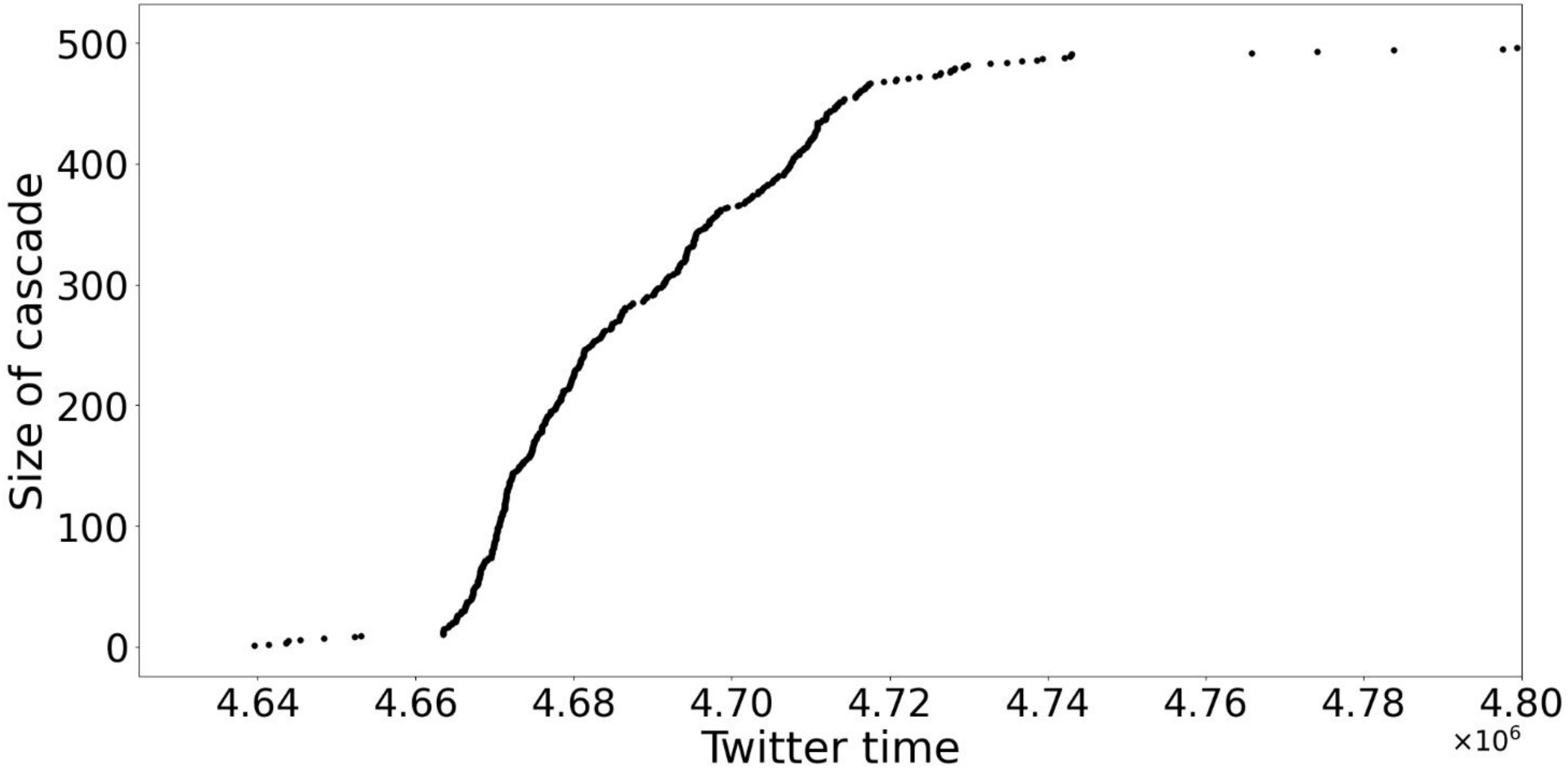


Number of tweets per hour
0
500
1000
1500
2000
2500
3000
3500
Example cascade (cumulative)
0
100
200
300
400
500
28-05
29-05
30-05
31-05
01-06
02-06
Time

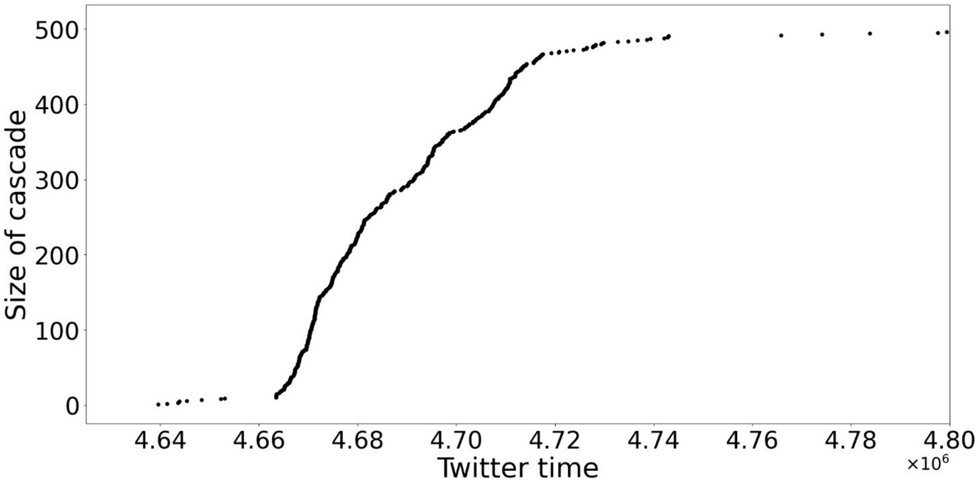

Cumulative number of tweets
175000
150000
125000
100000
75000
50000
25000
0
28-05
29-05
30-05
31-05
01-06
02-06
chronological time


Cumulative number of tweets
175000
150000
125000
100000
75000
50000
25000
0
4.62
4.65
4.67
4.70
4.72
4.75
4.78
4.80
×10^6
Twitter time

Twitter time



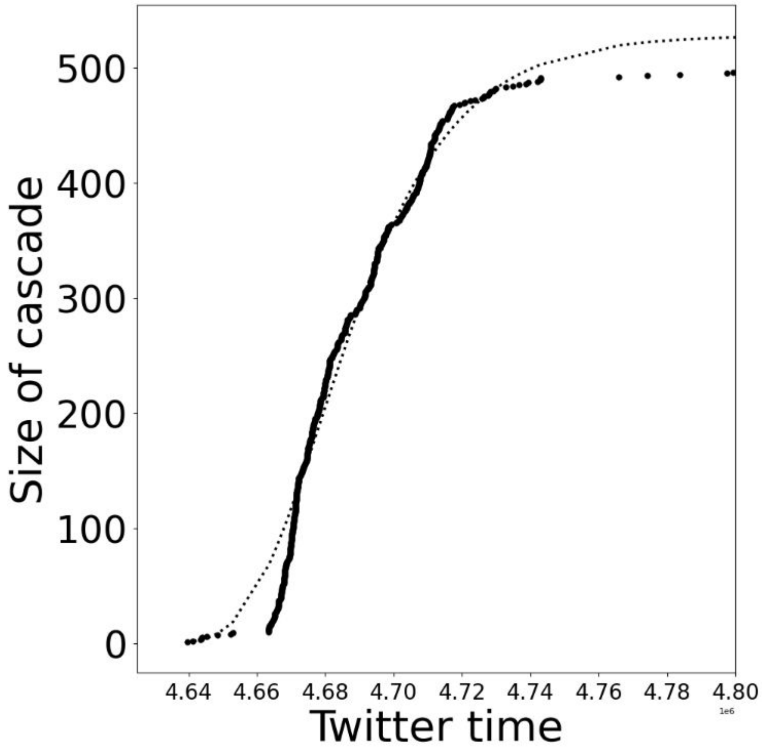
3-parameter model

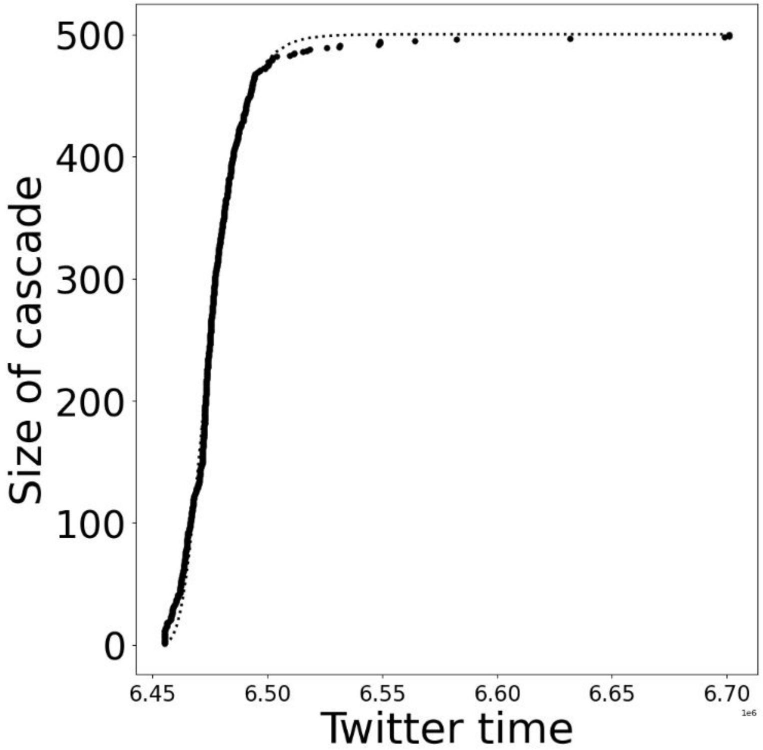


$$\frac{dN}{dt} = a \cdot e^{-g \cdot t}$$

$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

Model fits


Size of cascade
0
100
200
300
400
500
4.64
4.66
4.68
4.70
4.72
4.74
4.76
4.78
4.80
1e6
Twitter time


Size of cascade
0
100
200
300
400
500
6.45
6.50
6.55
6.60
6.65
6.70
1e6
Twitter time

3-parameter model: The size distribution

If we let all cascades grow to maximum size, (i.e., t→+∞), then:

$$\frac{dN}{dt} = a \cdot e^{-g \cdot t}$$

$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

$$\lim_{t \to \infty} N(t) = e^{a/g}$$

3-parameter model: The size distribution

If we let all cascades grow to maximum size, (i.e., $t \to +\infty$), then:

$$\frac{dN}{dt} = a \cdot e^{-g \cdot t}$$

$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

$$\lim_{t \to \infty} N(t) = e^{a/g}$$

Fitness

3-parameter model: The size distribution

If we let all cascades grow to maximum size, (i.e., t→+∞), then:

$$\frac{dN}{dt} = a \cdot e^{-g \cdot t}$$

$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

$$\lim_{t \to \infty} N(t) = e^{a/g}$$

Fitness, ω

What is the distribution of cascade **sizes (f(N))**, given a certain distribution of **fitnesses (g(ω))**?

$$f(N) = \frac{1}{N} \cdot g(\log N)$$

3-parameter model: The size distribution

If we let all cascades grow to maximum size, (i.e., t→+∞), then:

$$\frac{dN}{dt} = a \cdot e^{-g \cdot t}$$

$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

$$\lim_{t \to \infty} N(t) = e^{a/g}$$

Fitness, ω

What is the distribution of cascade **sizes (f(N))**, given a certain distribution of **fitnesses (g(ω))**?

$$f(N) = \frac{1}{N} \cdot g(\log N)$$

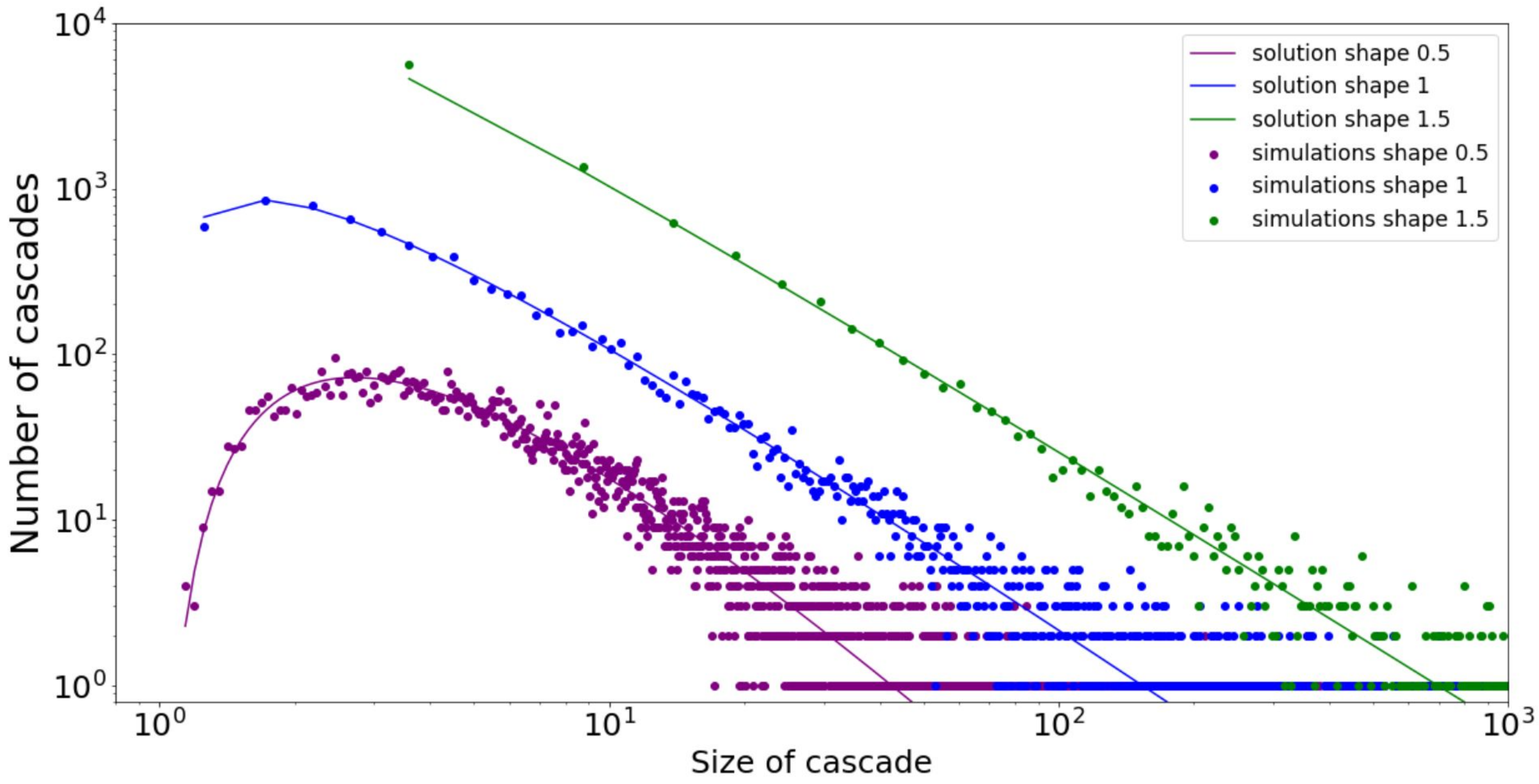
Example: **g(ω) is exponential**

$$g(\omega) = \lambda \cdot e^{-\lambda \cdot \omega}$$

$$f(N) = \frac{\lambda \cdot e^{-\lambda \cdot \log(N)}}{N}$$

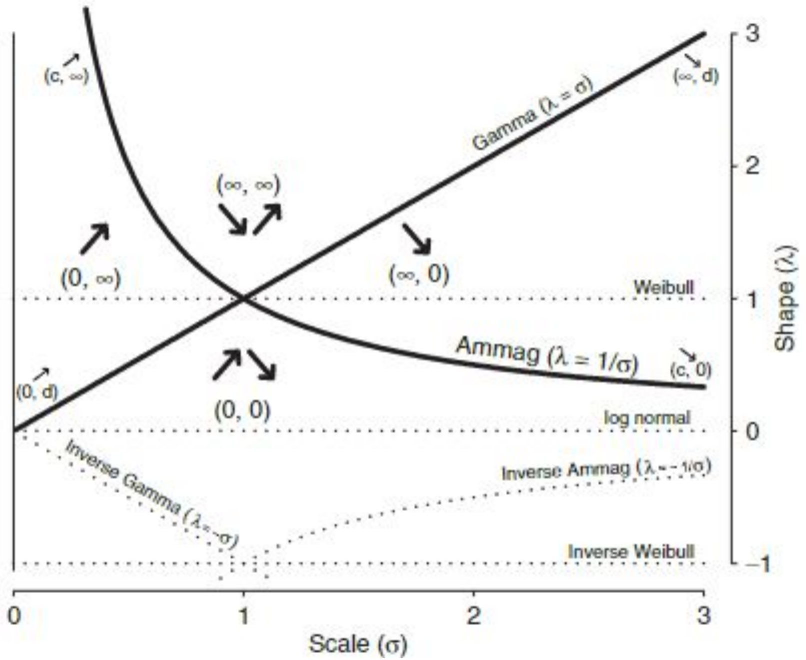
$$f(N) = \lambda \cdot N^{-\lambda - 1}$$

3-parameter model: The size distribution



Lília Perfeito

What is a good model for the fitness distribution?

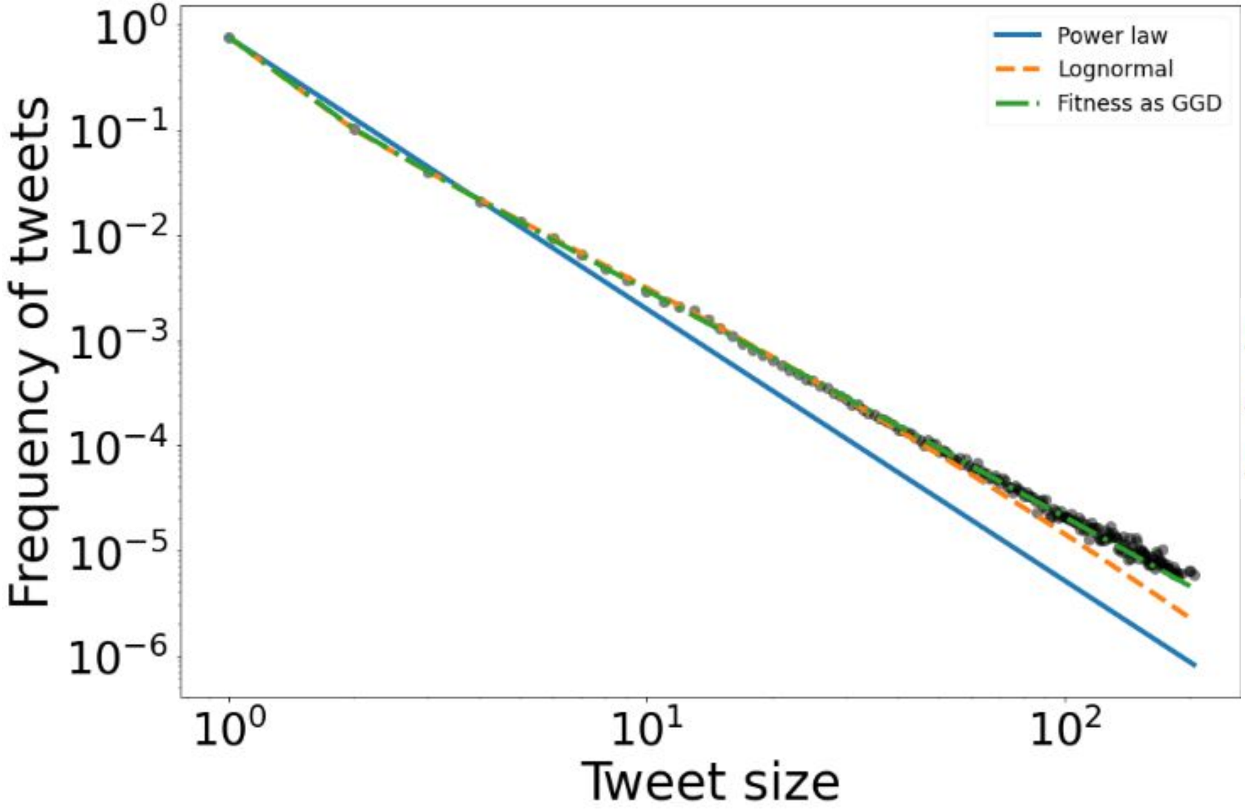


Exponential
Gamma
Lognormal
Powerlaw

Generalized Gamma

Cox, C., Chu, H., Schneider, M.F., Munoz, A.: Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in medicine **26**(23), 4352–4374 (2007)

Fitting cascade size distribution

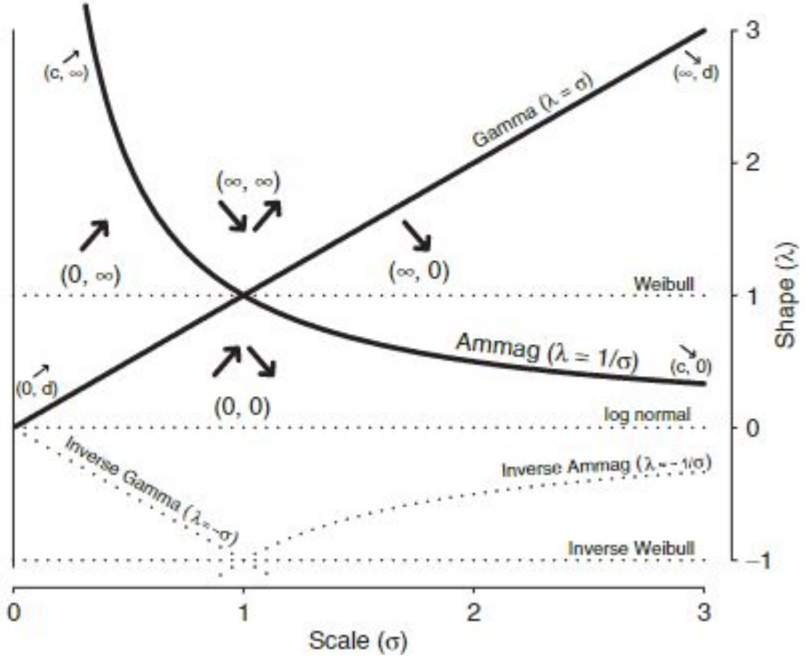


$$f(N) = a \cdot N^{-k}$$

$$f(N) = \text{lognormal}(N, \mu, \sigma \text{loc})$$

$$f(N) = \text{GGD}(\log(N), \sigma, \beta, \text{loc}) \cdot N^{-1}$$

Fitting cascade size distribution



		Parameters of fitness distribution			
Language	Dataset	λ	σ	β	location
Portuguese	Covid	0.23	1.28	20.58	-0.35
	Music	0.29	1.69	19.33	-0.21
	Film	0.29	1.57	18.11	-0.22
Italian	Covid	1.28	1.56	-0.20	-0.27
	Music	1.45	1.79	-0.51	-0.23
	Film	0.36	1.35	11.13	-0.31
German	Covid	1.96	1.53	-1.84	-0.23
	Music	2.22	2.05	-1.57	-0.15
	Film	1.98	2.05	-1.20	-0.17
Dutch	Covid	0.64	1.23	3.51	-0.39
	Music	0.76	1.26	2.80	-0.33
	Film	1.32	1.82	0.46	-0.21

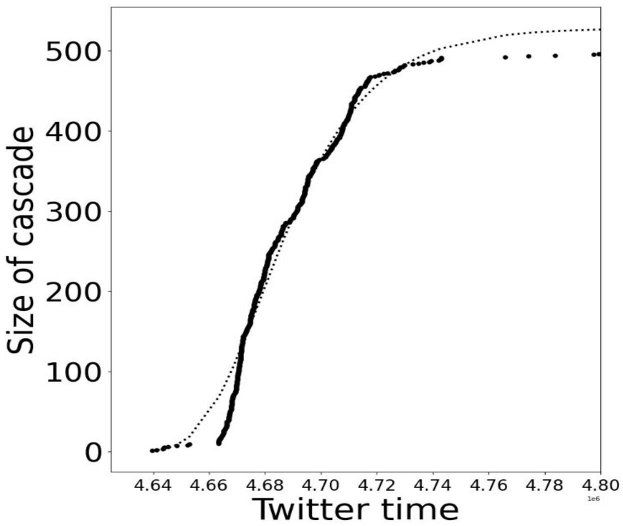
Intermediate conclusions

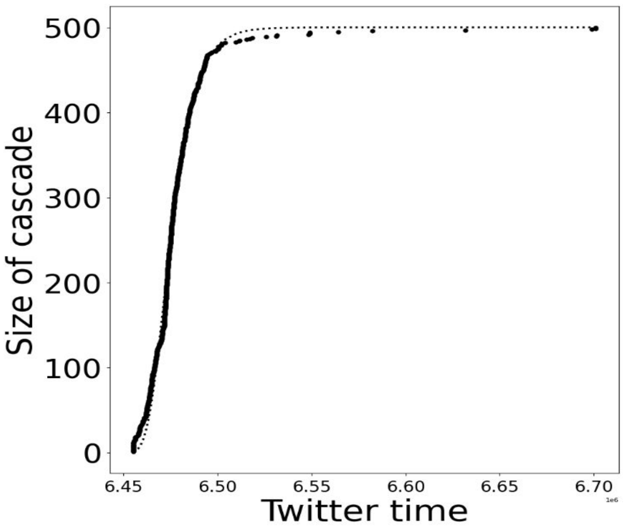
We introduce the concept of twitter time

We show the social network is not necessary to reproduce the cascade size distribution

The distribution of fitnesses is well approximated by a generalized Gamma distribution with an exponential-to-heavy tail

Dissecting the parameter space a little further

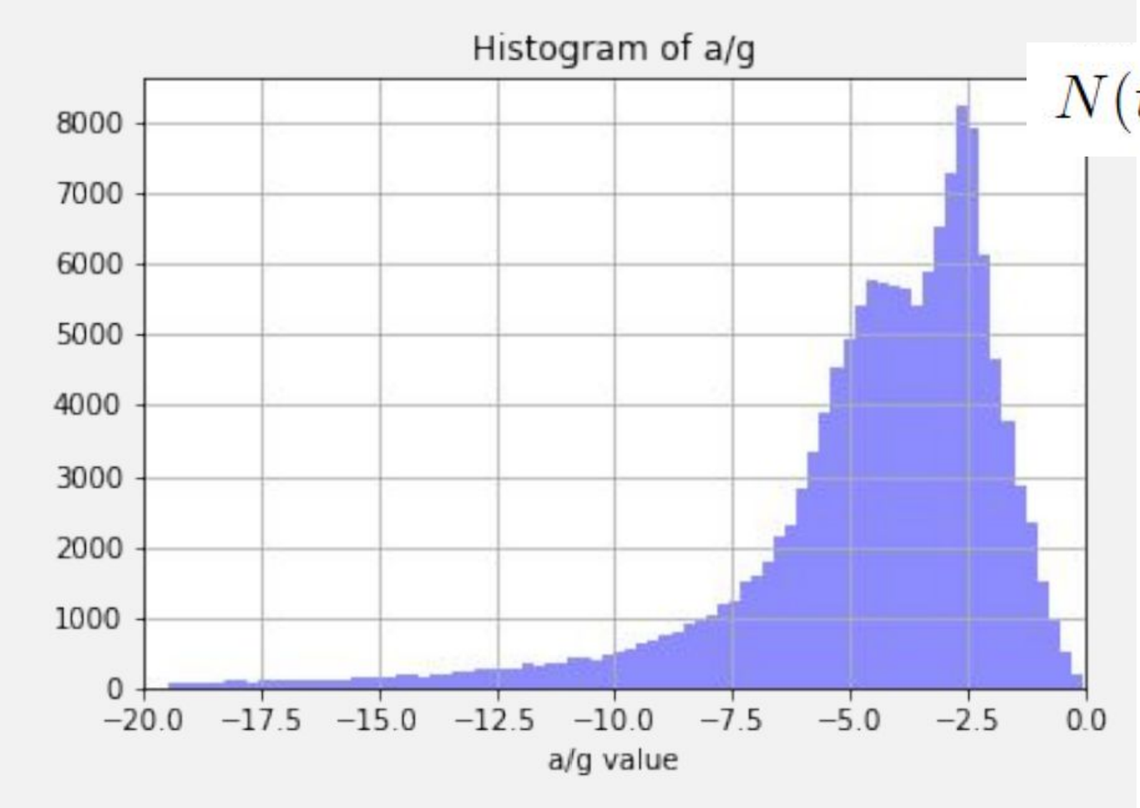




Another way to estimate the distribution of fitness parameters: fit each cascade 1 by 1

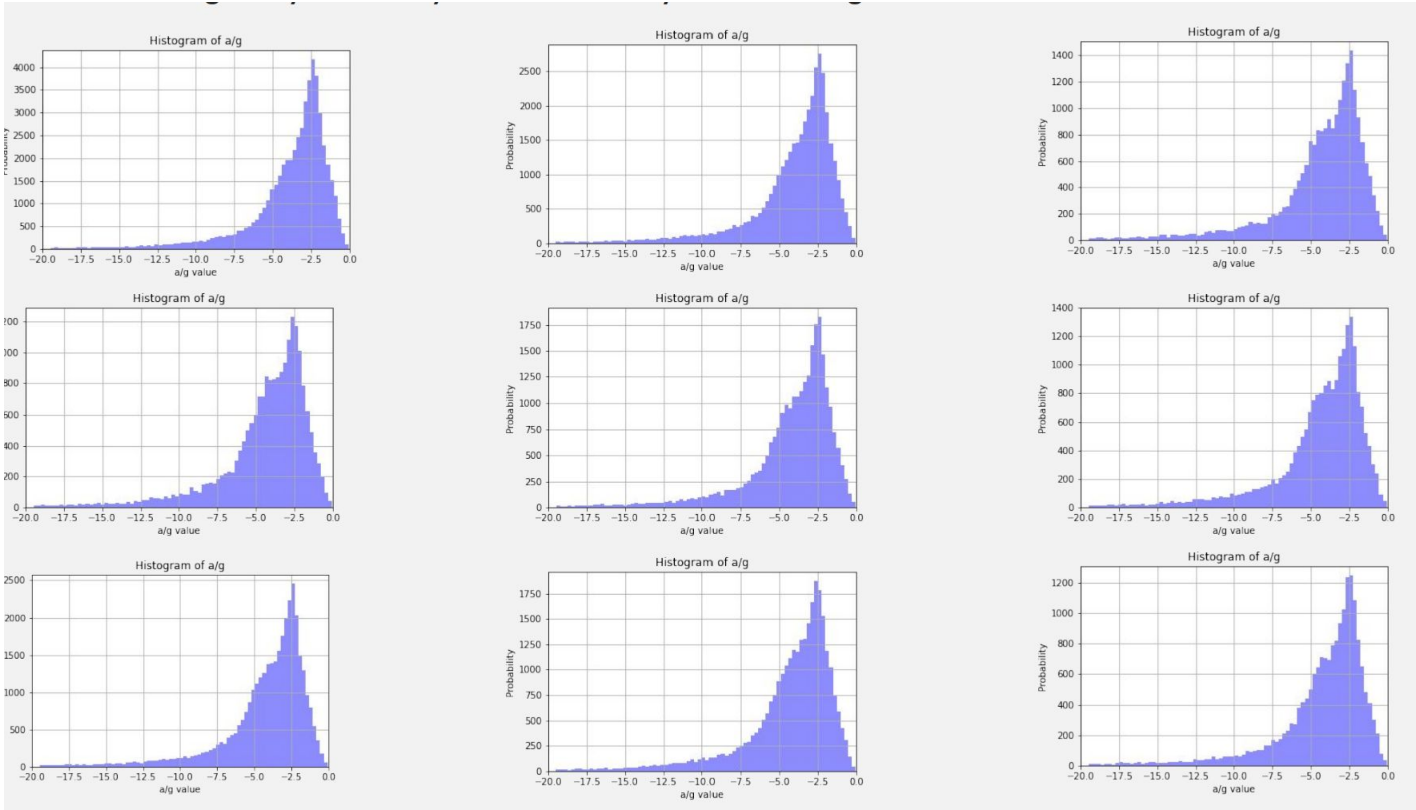
Time consuming, not applicable to all cases

Dissecting the parameter space a little further

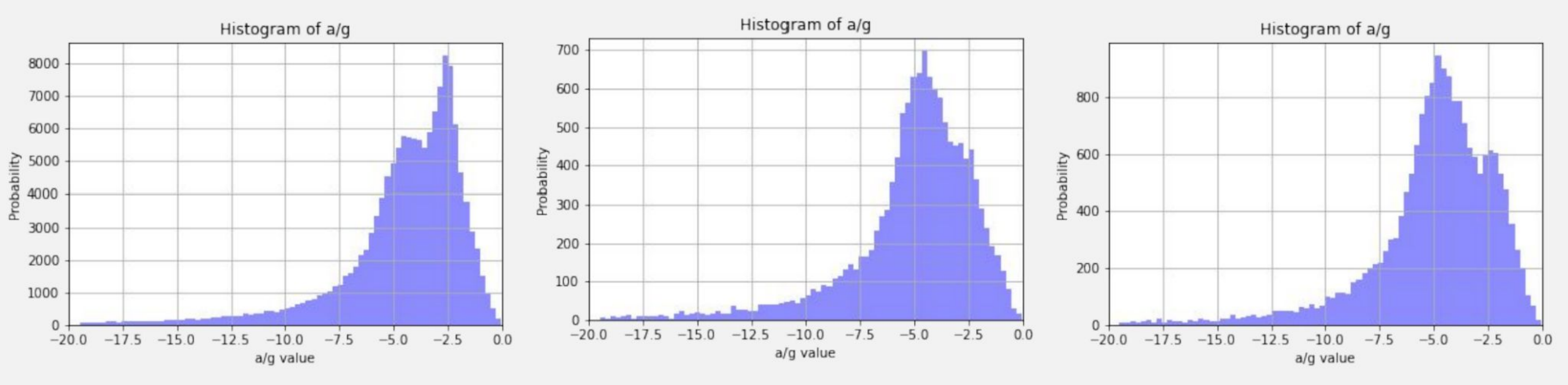


$$N(t) = N(0) \cdot e^{a/g} e^{-a \cdot e^{-g \cdot t}/g}$$

Dissecting the parameter space a little further

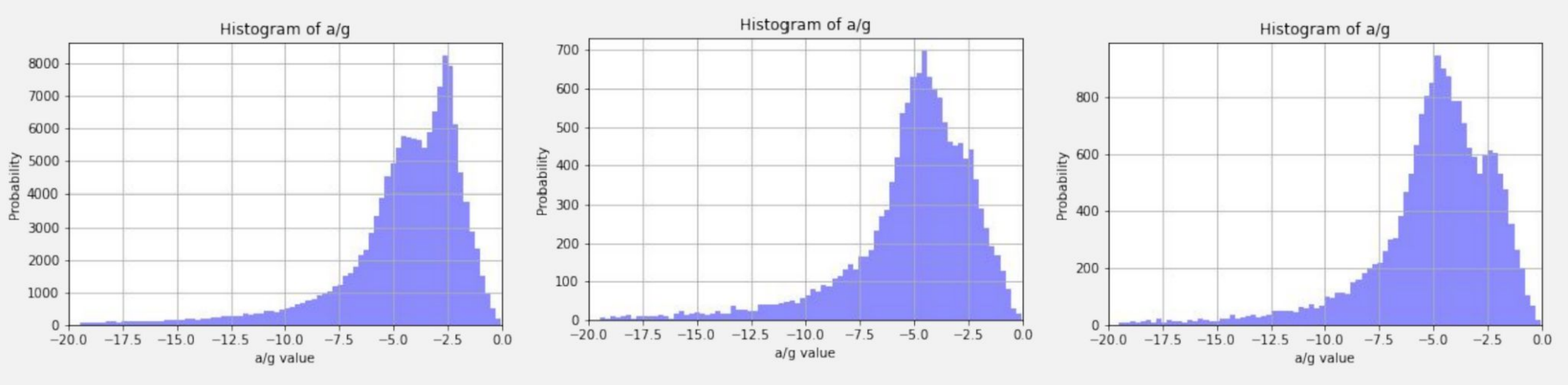


Dissecting the parameter space a little further



Covid **Film** **Music**

Dissecting the parameter space a little further



Covid **Film** **Music**

How can we separate them?

Can we tease apart the role of the network and growth?

Can we measure selection acting on information?

Back to the questions

Q1: Can we tease apart the role of the network and growth?

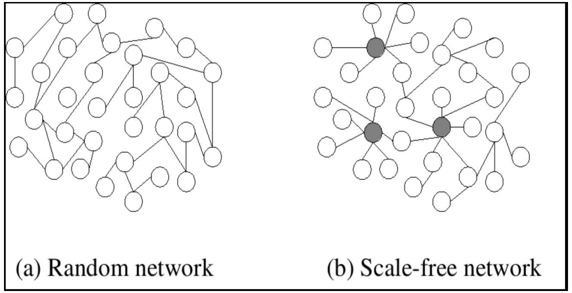
A1: The network is not necessary.

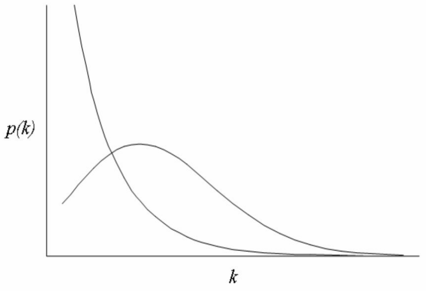
Q1.1: Which one explains more of the variability in the data?

Q2: Can we measure selection acting on information?

A2: Yes

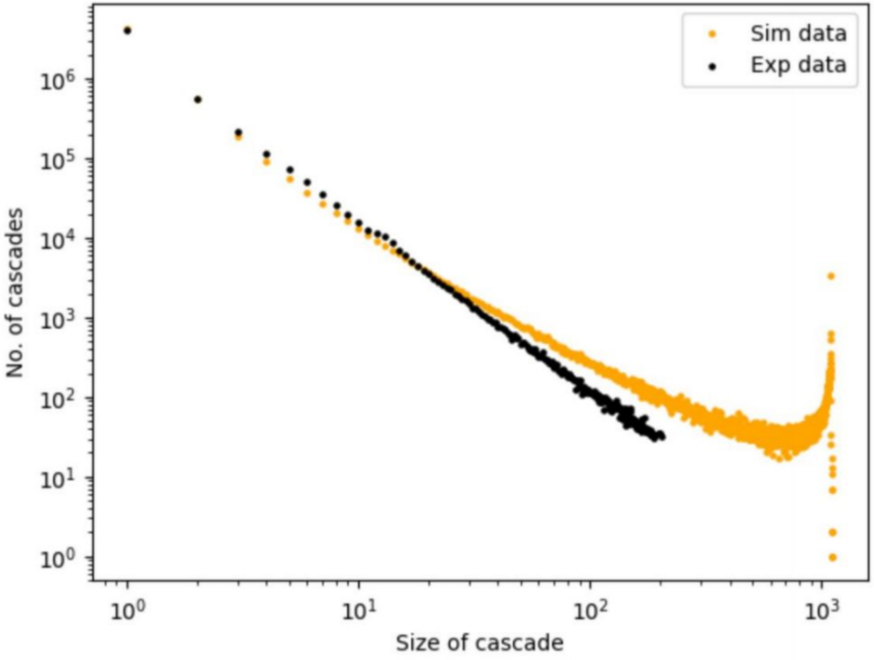
Simulating tweets spreading on a social network

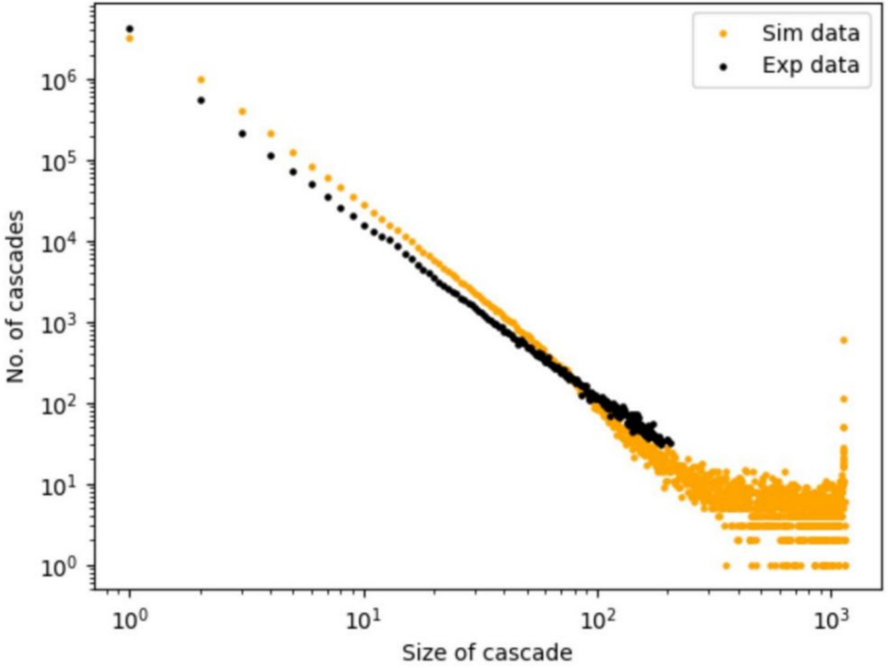




1. Build the network (1000 nodes)
   a. Uniform/Poisson
   b. Realistic, ie, powerlaw
2. Choose a node at random to post the tweet
3. Let tweets spread
   a. Following $p_i(\text{share}) = k_i \cdot a_c \cdot e^{-g_c \cdot t}$
   b. Introducing a set of resistant nodes
4. Repeat thousands of times, find the simulation parameters that best fit the data (Approximate Bayesian Computation)
5. Compare the goodness of fit of the best model without network with the best model with network

First simulation results





Outlook

- Simple *vs* complex contagion
- Introducing heterogeneity in nodes (personality types)
- True *vs* misinformation
- …