Supplementary Materials: Contagion dynamics in real-life and digital settings

Adam J. Kucharski[1*]

[1]Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, UK.
*To whom correspondence should be addressed. Email: adam.kucharski@lshtm.ac.uk.

1 Data and parameter estimates

This section details the data and methods used to obtain the illustrative estimates shown in Figure 2. All estimates are based on data described in the main text of published studies, or extracted from accompanying figures and tables. All code and data required to reproduce Figures 2 and 3 are available from:

https://github.com/adamkucharski/contagion-review

1.1 Reproduction number and serial interval

1.1.1 Burglary

The spread of burglary in Los Angeles was inferred using a Hawkes process applied to a dataset of 5,376 reported residential burglaries in San Fernando Valley collected the Los Angeles Police Department during 2004–2005 (*1*), under the assumption that contagion was present. 13.1% of all events were estimated to be offspring of previous events, which implies each non-offspring event generated an outbreak of size 1/0.131=1.15, resulting in an estimate of $R = 0.131$. There were two peaks in the estimate of time between events: 1–2 days (which the authors call 'fast' crime) and 7 days ('routine' crime). The highest peak was the shorter delay, so a serial of interval of 1.5 days is shown in Figure 2.

1.1.2 Ebola

The basic reproduction number (i.e. R in a fully susceptible population) estimated up to 6th July 2014 in Sierra Leone during the 2013–16 West Africa Ebola epidemic was

2.02 (95% CI: 1.79–2.26) (*2*). The serial interval across all affected countries (Liberia, Nigeria, Sierra Leone, Guinea) up to 14th September 2014 was 15.3 (SD = 9.1).

1.1.3 Email chains

Data included electronic communications between 8,952 volunteer employees within a large company, based in over 70 countries over a two year period (*3*). Median time for e-mail forwarding (i.e. serial interval) ranged from 7–18 depending on job roles (values extracted from Figure 6 in (*3*)), with an overall median of 10.5 hours. R was estimated from the cascade size distribution (extracted from in Figure 10). The mean cascade size was 4, which translates into R=0.75.

1.1.4 Email marketing

Data were from viral marketing campaigns in eleven European markets; these offered incentives to people who subscribed to the newsletter of an online IT company online newsletter if they recommended the subscription to others via e-mail; in total, 7,225 people initiated diffusion cascades (*4, 5*). Across all markets, the analysis produced an estimate of R=0.246, with participants taking an average of 1.5 days to forward the message (SD = 5.5 days).

1.1.5 Feynman diagrams

Fitting transmission dynamic rumour models to time series data on the cumulative number of authors using Feynman diagrams between 1948 and 1960, the mean R was estimated to be 14.975 (SD = 2.227) in the US, 65.245 (SD = 13.808) in Japan, and 25.055 (SD = 10.614) in Japan (*6*). The mean rates of loss of infectiousness was 0.049 (SD = 0.034), 0.054 (SD = 0.034), and 0.092 (SD = 0.062) for the three countries respectively. The duration of infectiousness was therefore 20.4 years in the US, 18.5 years in Japan and 10.9 years in USSR. If infection prevalence is at equilibirum, and the incubation and latent periods follow the same distribution, this corresponds to the serial interval of infection.

1.1.6 Gun violence

Transmission chains of shootings between co-offenders arrested in Chicago between January 1, 2006, and March 31, 2014 were inferred using a Hawkes process (*7*). The mean size of the inferred transmission clusters was 2.3, resulting in an estimate of R=0.63. The

serial interval (i.e. mean time from one shooting to the next in the transmission chains) was 125 days.

1.1.7 Ice bucket challenge

Based on observed chains of nominations between 91 celebrities in 2014 (*8*), R was estimated to be 1.43 (95% CI: 1.23–1.65) with a serial interval of 2.1 days.

1.1.8 Influenza A/H1N1p

The basic reproduction number (i.e. R in a fully susceptible population) estimated from infections in Mexico up to the end of April 2009 1.58 (95% CI: 1.34?2.04) (*9*). The generation time (i.e. time from infection of infector to infection of infectee, which equivalent to the serial interval if the incubation and latent period follow the same distribution) was 1.91 days (95% CI: 1.30–2.71).

1.1.9 Influenza A/H5N1

Based on a study of influenza A/H5N1 in households in Indonesia from 2005–2009 (*10*), the mean serial interval was 5.6 days, with R estimated to be 0.140 (95% CI: 0.004–0.390).

1.1.10 Laughter

A study of laughter contagion examined outcomes among college students following exposure to an audio 'laughter box' (*11*). The study included three groups, with the probability of laughter calculated for each: n=49, p=0.57; n=18, p=0.66; n=61, p=0.33. Hence the overall probability of transmission was 0.47. This is equivalent to R for a one-on-one encounter. The study suggested that 'the latency of the contagious response to canned laughter seems to be almost immediate'; for visualisation purposes, the serial interval was therefore assumed to be 5 seconds (the second fastest serial interval shown in Figure 2 is 22 seconds (*12*), so this assumption does not affect the overall ordering).

1.1.11 LaTeX macros

Transmission chains were directly calculated by tracking LaTeX macros in arXiv publications (*13*). The mean time between adoptions (i.e. serial interval) for macro X (defined

below) was 5.5 years; the serial interval for macro Y (defined below) was 2.8 years. As the study used data up to November 2015, and the larger serial interval was 5.5 years, adoptions in the final six years of the dataset were assumed to still be capable of producing as-yet-unidentified transmission events. Note that there were gaps in publications between May 2002–April 2014, and November 2006–January 2010 in the two respective datasets, which makes this cutoff less arbitrary.

Between February 1993 and April 2015, a total of 102 authors adopted the macro X. Subtracting adoptions in the final six years, 75 authors in total had used the macro. Hence $R = 102/75 = 1.36$. Between June 2001 and February 2014, a total of 61 authors adopted the macro Y. Subtracting adoptions in the final six years, 48 authors in total had used the macro. Hence $R = 61/48 = 1.27$.

Macro X: `\mathrel{\mathpalette\@versim>}`

Macro Y: `\mbox{\boldmath $Y$}`

1.1.12 Mass shootings and school shootings

The reproduction numbers and serial intervals for contagion of mass shootings in the US between 2006 and December 2013 and school shootings in the US between 1997 and 2013 was estimated using a Hawkes process (*14*). The model estimated R=0.3 (95% CI: 0.12–0.56) and serial interval 13.2 days (95% CI: 4.3–46.8) for mass shootings, and R=0.22 (95% CI: 0.10–0.42) and serial interval 12.9 days (95% CI: 5.4–53.3) for school shootings.

1.1.13 MERS-CoV

Based on clusters of MERS-CoV reported in the Kingdom of Saudi Arabia between January 2013 and July 2014 (*15*), the mean serial interval was 6.8 days (95% CI: 6.0–7.8), with the overall reproduction number estimated to be 0.45 within clusters, 0.24 within regions and 0.05 between regions. Hence the overall value of R was 0.74.

1.1.14 Monkeypox

Based on the size distribution of transmission clusters reported in Democratic Republic of Congo from 1980–1984, which had a mean size of 1.42, the estimate for R=0.3 (*16*). The mean serial interval between the 47 cases who were classified as secondary and their infectors was 13.8 days (*17*).

1.1.15 Popular Facebook content

Analysis was based on 98 high profile sharing cascades on Facebook from mid-2014 to early 2016 (*12*). There were four types of content: transient (i.e. content visible for short period of time, like an item shared on a newsfeed); persistent/pinned (i.e. user changing profile picture to support a cause); volunteer-based (i.e. user invites friends to respond to a post, such as a word play meme, friends sign up and are allocated tasks, then they share the meme with their own friends once tasks are complete); and nomination-based (i.e. user completes an activity then nominates others to do the same, such as the Ice Bucket Challenge). The study recorded the secondary sharing (i.e. R) as well as delay between more recent exposure and sharing, which can be viewed as a lower bound for the serial interval. Based on reconstructed transmission chains, transient content had R=1.84 and delay between posts (i.e. lower bound serial interval) of 23 seconds; persistent had R=2.56 and interval of 153s; volunteer had R=2.56 and interval of 1.31×10^4s; and nomination had R=1.76 and interval of 4.42×10^4s.

1.1.16 SARS-CoV

The basic reproduction number (i.e. R in a fully susceptible population) estimated based on data from the first ten weeks of the 2003 outbreak in Hong Kong was 2.7 (95% CI: 2.2–3.7) (*18*). The serial interval based on the first 205 probable cases of SARS reported in Singapore in Singapore was 8.4 days (SD = 3.8) (*19*).

1.1.17 Scientific fields

The same modelling approached described in section was applied to time series data for different scientific fields, specifically the cumulative number of authors publishing (*20*). The serial interval τ was derived in the same way, from the estimated duration of infection. The fields and estimates were as follows: cosmological inflation ($R = 64 \pm 1.5$, τ = 4.76 years); cosmic strings ($R = 2.58 \pm 0.11$, τ = 0.58 years); prions ($R = 1.87 \pm 0.03$, τ = 2.70 years); H5N1 influenza research ($R = 2.44 \pm 0.03$, τ = 1.67 years); carbon nanotubes ($R = 9.72 \pm 1.71$, τ = 10.00 years); quantum computing ($R = 3.2 \pm 0.11$, τ = 0.85 years).

1.1.18 Self-harm in adolescent units

A study in an adolescent psychiatric ward in Finland identified multiple clusters of self-harm over a 12 month period (*21*). These were of size: 1, 1, 2, 2, 2, 3, 3, 5. A contagion event between two individuals was defined as one that occurred either on the same day or the consecutive day of another. Focusing on contagion at the overall person-level, rather than considering repeat incidents between the same two individuals, the mean cluster size across these studies was 4.8, which would suggest $R = 1 - 1/2.38 = 0.58$. Based on the definition of contagion in these studies, the serial interval was assumed to be one day for visualisation purposes in Figure 2.

Another study in an adolescent psychiatric unit used detailed case investigations to reconstruct the path of a 1964 outbreak of self-injury (*22*). Overall 11 individuals were involved, suggesting $R = 1 - 1/11 = 0.91$. Based on the reconstructed path of direct influence between individuals (Figure 2 in (*22*)), the serial interval was around 7 weeks.

1.1.19 Smallpox

Based on a transmission model applied to data from the 1967 outbreak in Abakaliki, Nigeria, the estimated value of R_0 (i.e. R in a fully susceptible population) was 6.87 (95%CI: 4.52–10.1) (*23*). A systematic review found that smallpox had a mean serial interval of 17.7 days (*24*).

1.1.20 Twitter cascades

Based on retweet chains on Twitter between 1st July 2013 and 31st July 2014, there were 0.87 retweets per tweet on average, i.e. $R = 0.87$ (*25*). The distribution of delays from one tweet to another for tweets with neutral sentiment had a mean of 7.3 hours.

1.1.21 Yawns

In a study of yawn contagion among pairs of adults (*26*), the estimated mean probability of yawn occurrence following a trigger yawn was around 0.08 when the pair were strangers and 0.51 when they were kin. Hence between a pair of strangers, we would expect one yawn to trigger around R=0.08 secondary yawns, and between kin, R=0.51 yawns. Based on the estimated delays between yawns, there was an expected delay (i.e. serial interval) of around 1.12 minutes for kin and 2.14 minutes for strangers.

1.2 Doubling time

1.2.1 Code Red worm

The Code Red worm starting spreading via the internet on 19th July 2001. The number of infected hosts was estimated to have doubled in size every 37 minutes during the early stages of the oubtreak (*27*).

1.2.2 Ebola

See section 1.1.2 below for full details of data (*2*). The doubling time estimated up to 6th July 2014 in Sierra Leone was 12.84 days (95% CI: 10.92–15.66).

1.2.3 Feynman diagrams

See section 1.1.17 below for full details of data (*6*). The doubling time of each type of content was calculated as $\tau / \log_2(R)$, where τ is the serial interval and R is the reproduction number.

1.2.4 Influenza H1N1p

See section 1.1.8 below for full details of data (*9*). Based on genetic data, the doubling time of the epidemic was estimated to be 10 days (95% CrI: 4.5—37.5)

1.2.5 Popular Facebook content

See section 1.1.15 below for full details of data (*12*). The doubling time of each type of content was calculated as $\tau / \log_2(R)$, where τ is the serial interval and R is the reproduction number.

1.2.6 Slammer worm

The Slammer worm starting spreading via the internet on 25th January 2002. The number of infected hosts was estimated to have doubled in size every 8.5 seconds during the early stages of the oubtreak (*27*).

1.2.7 Stamps

The number of European countries using postage stamps grew exponentially from 1 to 24 between 1840 and 1860 (*28*). This corresponds to a doubling time of $20 / \log_2(1/24) =$

4.36 years.

1.2.8 VCRs

The percentage of US households that owned VCRs increased from 2.2% in 1981 to 18% in 1985 (*29*). This corresponds to a doubling time of $4/\log_2(2.2/18) = 1.32$ years .

1.2.9 WannaCry randomware

The WannaCry ransomeware spread widely online during May 2017. On 12th May, the number of exploit attempts of the Windows vulnerability used by WannaCry that were blocked by Symantec grew exponentially from around 20,000 to 80,000 in the space of 7 hours (*30*). This corresponds to a doubling time of $7/\log_2(8/2) = 1.33$ hours.

2 Literature review

A search was conducted on Web of Science to identify studies of diffusion on different social media platforms, published from 2009–2018 (as the focus was on recent levels of research activity). The search strategy was refined to account for differing terminology used in studies of the spread of online content.

The final search string used was:

(<platform name> AND (contagio* OR diffus* OR transmi* OR propagat*))

Title, abstract, and where necessary full text, were reviewed to obtain a set of relevant studies for each platform.

Publications matching the following criteria were included:

- Studies that looked at the diffusion of content, information, or behaviour via that specific platform.
- Mathematical modelling studies of diffusion that were directly informed by that specific platform.

Publications matching the following criteria were excluded:

- Studies that only mentioned the platform as an illustrative or comparative example (e.g. 'Weibo is a Twitter-like platform' would be excluded as a study of Twitter).
- Studies focusing on adoption of the platform itself in a population, rather than diffusion via the platform.
- Technical papers discussing streaming, compression, or transmission (i.e. broadcasting) of multimedia content.

A summary of the search is shown in Table S1, with final results in Table S2.

	Twitter	Facebook	Instagram	YouTube	Reddit	WhatsApp	Weibo
WoS search	1165	618	33	216	16	45	205
After screening	831	231	11	78	10	4	156

Table S1: Number of papers in initial Web of Science search and after screening.

Year	Twitter	Facebook	Instagram	YouTube	Reddit	WhatsApp	Weibo
2009	3	1	0	3	0	0	0
2010	4	5	0	2	0	0	0
2011	15	8	0	3	0	0	2
2012	44	15	0	6	0	0	5
2013	51	35	0	4	0	0	10
2014	108	25	1	7	1	0	24
2015	125	29	0	11	2	0	29
2016	145	40	0	13	0	0	31
2017	169	37	1	12	2	2	36
2018	167	36	9	17	5	2	19
Users (m)	330	2300	1000	1800	1600	1500	431

Table S2: Number Web of Science papers identified for each platform after screening. Also shown are approximate monthly active users in 2018 (in millions) for each platform, apart from YouTube, where activity is measured by worldwide viewers (*31–36*).

References

1. G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, G. E. Tita, *Journal of the American Statistical Association* **106**, 100 (2011).

2. WHO Ebola Response Team, *New England Journal of Medicine* **371**, 1481 (2014).

3. D. Wang, *et al.*, *Proceedings of the 20th international conference on World wide web* (ACM, 2011), pp. 735–744.

4. J. L. Iribarren, E. M. Egido, *PhD Thesis* (2015).

5. J. L. Iribarren, E. Moro, *Social networks* **33**, 134 (2011).

6. L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, C. Castillo-Chávez, *Physica A: Statistical Mechanics and its Applications* **364**, 513 (2006).

7. B. Green, T. Horel, A. V. Papachristos, *JAMA Internal Medicine* **177**, 326 (2017).

8. M. Y. Ni, B. H. Y. Chan, G. M. Leung, E. H. Y. Lau, H. Pang, *BMJ* **349**, g7185 (2014).

9. C. Fraser, *et al.*, *Science* **324**, 1557 (2009).

10. T. Y. Aditama, *et al.*, *PLoS ONE* **7**, e29971 (2012).

11. R. R. Provine, *Bulletin of the Psychonomic Society* **30**, 1 (1992).

12. J. Cheng, *et al.*, *Twelfth International AAAI Conference on Web and Social Media* (2018).

13. R. Rotabi, C. Danescu-Niculescu-Mizil, J. Kleinberg, *Eleventh International AAAI Conference on Web and Social Media* (2017).

14. S. Towers, A. Gomez-Lievano, M. Khan, A. Mubayi, C. Castillo-Chavez, *PLOS ONE* **10**, e0117259 (2015).

15. M. Gabielkov, A. Ramachandran, A. Chaintreau, A. Legout, *ACM SIGMETRICS Performance Evaluation Review* **44**, 179 (2016).

16. S. Blumberg, J. O. Lloyd-Smith, *PLoS Comput Biol* **9**, e1002993 (2013).

17. P. Fine, Z. Jezek, B. Grab, H. Dixon, *International journal of epidemiology* **17**, 643 (1988).

18. S. Riley, *et al.*, *Science* **300**, 1961 (2003).

19. M. Lipsitch, *et al.*, *Science* **300**, 1966 (2003).

20. L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, C. Castillo-Chávez, D. E. Wojick, *Scientometrics* **75**, 495 (2008).

21. T. J. Taiminen, K. Kallio-Soukainen, H. Nokso-Koivisto, A. Kaljonen, H. Helenius, *Journal of the American Academy of Child & Adolescent Psychiatry* **37**, 211 (1998).

22. P. Matthews, *International Journal of Social Psychiatry* **14**, 125 (1968).

23. M. Eichner, *American Journal of Epidemiology* **158**, 110 (2003).

24. M. A. Vink, M. C. J. Bootsma, J. Wallinga, *American journal of epidemiology* **180**, 865 (2014).

25. S. Tsugawa, H. Ohsaki, *Social Network Analysis and Mining* **7** (2017).

26. I. Norscia, E. Palagi, *PLOS ONE* **6** (2011).

27. D. Moore, *et al.*, The spread of the sapphire/slammer worm, *Tech. rep.*, CAIDA, ICSI, Silicon Defense, UC Berkeley EECS and UC San Diego CSE (2003).

28. H. E. Pemberton, *American Sociological Review* **1**, 547 (1936).

29. J. Romaine, *RFID Journal* (2015).

30. SSL Support Desk, *https://www.sslsupportdesk.com/wannacry-blocked-by-symantec-best-practices-against-ransomware/* (2017).

31. Statista, *https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/* (2018).

32. Statista, *https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/* (2018).

33. Statista, *https://www.statista.com/statistics/805656/number-youtube-viewers-worldwide/* (2018).

34. Statista, *https://www.statista.com/statistics/443332/reddit-monthly-visitors/* (2018).

35. Statista, *https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/* (2018).

36. China Internet Watch, *https://www.chinainternetwatch.com/26225/weibo-q2-2018/* (2018).