ENCISE
ENhancing seCurity and privAcy in the Social wEb

IDRAMA LAB

Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro,
Michael Sirivianos, Gianluca Stringhini, Jeremy Blackburn

Cyprus University of Technology

UCL

BOSTON UNIVERSITY

UAB

THE**RUSSIA**INVESTIGATION

How Russian trolls manipulated American politics

By Marshall Cohen, CNN

Updated 0220 GMT (1020 HKT) October 20, 2018

ARGUMENT

How Russia Sows Confusion in the U.S. Vaccine Debate

Not content to cause political problems, Moscow's trolls are also undermining public health.

BY KATHERINE KIRK | APRIL 9, 2019, 2:48 PM

SOCIAL TWITTER POLITICS

Twitter's list of 2,752 Russian trolls

From @10_gop to @ZzzacharyZzz.

By Dan Frommer | @fromedome | Nov 2, 2017, 11:45am EDT

Research Questions

- How do state-sponsored actors operate and evolve?
- How does the behavior of state-sponsored trolls compare to random users?
- More importantly, what was their influence on the Web with respect to the dissemination of news?
  - Focus on Twitter, Reddit, 4chan's /pol/

Datasets

- Russian trolls dataset
  - Look for tweets from the 2.7K identified troll accounts
  - 27K tweets from 1K identified troll accounts
- Random dataset
  - Extract a set of 1K users that have “similar” posting activity as the Russian trolls
  - 96K tweets from 1K random users
- Influence Estimation Datasets
  - Twitter 1% Streaming API dataset
  - 4chan’s /pol/ posts from Hine et al. (ICWSM’17)
  - Reddit submission and comments from Pushshift

Results

What are they posing as?

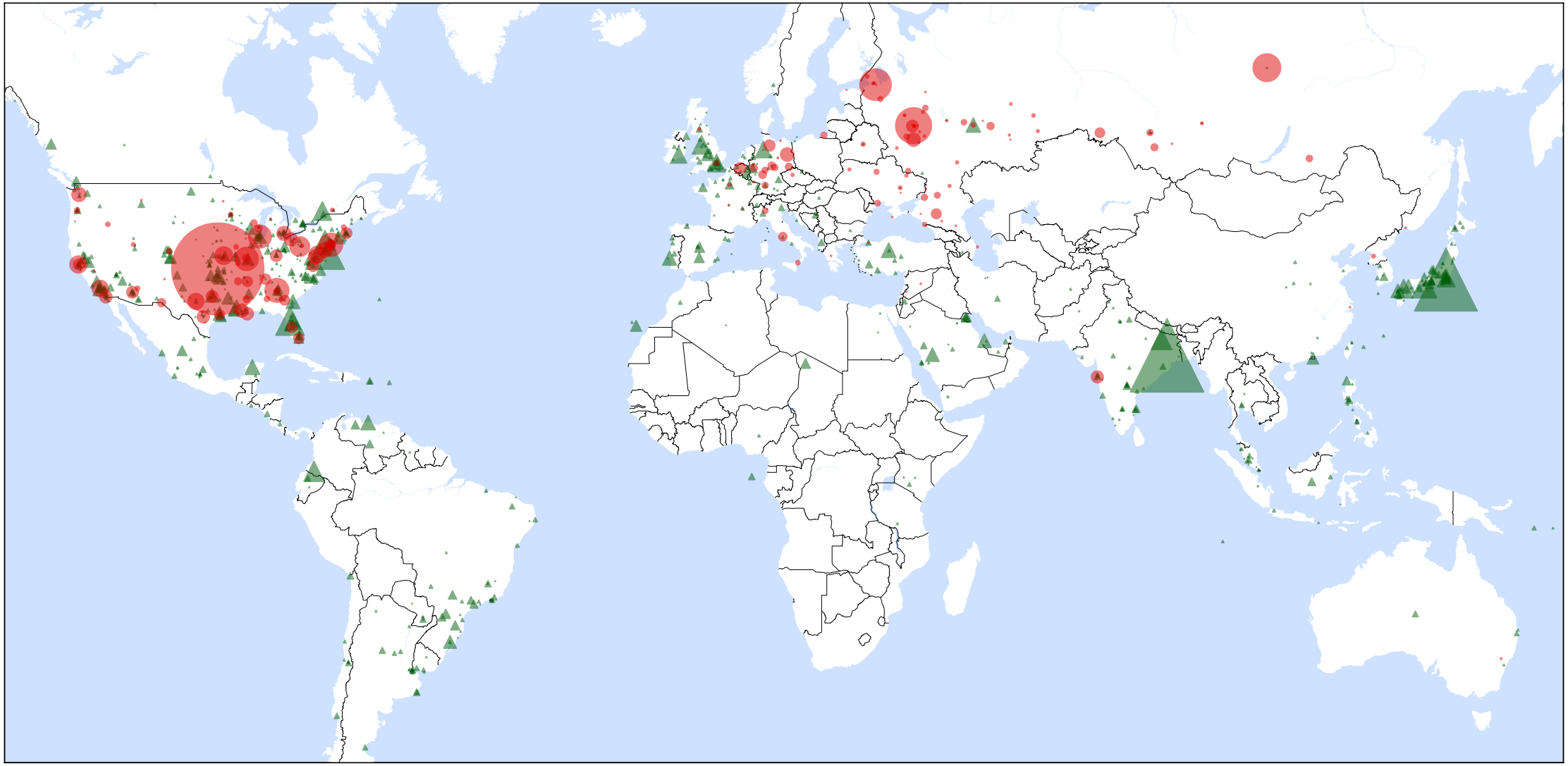
News accounts?

Nudging users to follow them

Trump supporters

Description			
Word	**(%)**	**Word bigram**	**(%)**
news	10.7%	follow me	7.8%
follow	10.7%	breaking news	2.6%
conservative	8.1%	news aus	2.1%
trump	7.8%	uns in	2.1%
und	6.2%	deiner stdt	2.1%
maga	5.9%	die news	2.1%
love	5.8%	wichtige und	2.1%
us	5.3%	nachrichten aus	2.1%
die	5.0%	aus deiner	2.1%
nachrichten	4.3%	die dn	2.1%

Where are they allegedly posting from?



What Twitter clients did trolls used?

Mostly through browser

Dashboard

Mobile apps

Client (Trolls)	(%)	Client (Baseline)	(%)
Twitter Web Client	50.1%	TweetDeck	32.6%
twitterfeed	13.4%	Twitter for iPhone	26.2%
Twibble.io	9.0%	Twitter for Android	22.6%
IFTTT	8.6%	Twitter Web Client	6.1%
TweetDeck	8.3%	GrabInbox	2.0%
NovaPress	4.6%	Twitter for iPad	1.4%
dlvr.it	2.3%	IFTTT	1.0%
Twitter for iPhone	0.8%	twittbot.net	0.9%
Zapier.com	0.6%	Twitter for BlackBerry	0.6%
Twitter for Android	0.6%	Mobile Web (M2)	0.4%

These days Twitter reports client information on each tweet

What hashtags they shared?

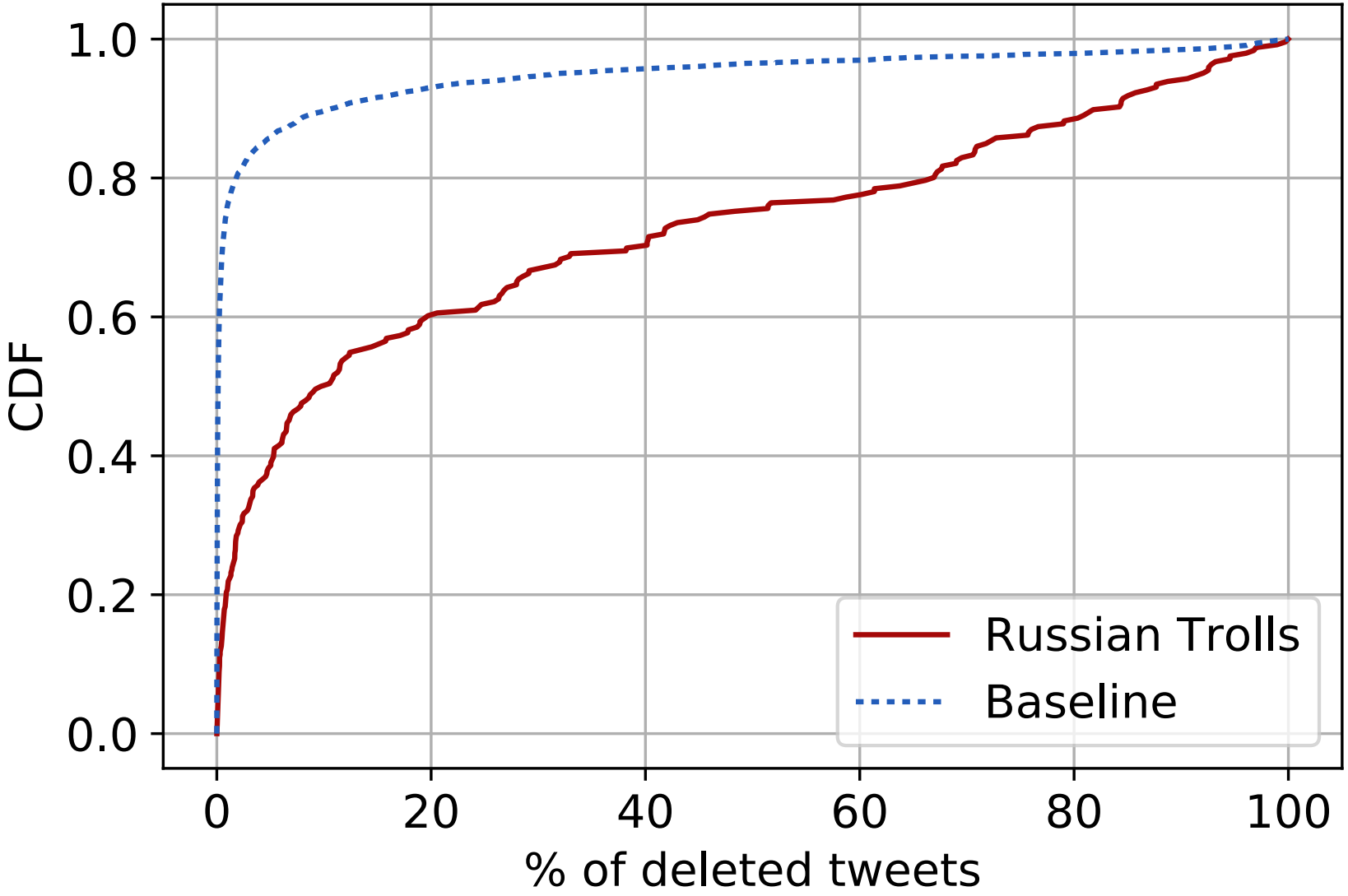
Trolls				Baseline			
Hashtag	(%)	Hashtag	(%)	Hashtag	(%)	Hashtag	(%)
news	7.2%	US	0.7%	iHeartAwards	1.8%	UrbanAttires	0.6%
politics	2.6%	tcot	0.6%	BestFanArmy	1.6%	Vacature	0.6%
sports	2.1%	PJNET	0.6%	Harmonizers	1.0%	mPlusPlaces	0.6%
business	1.4%	entertainment	0.5%	iOSApp	0.9%	job	0.5%
money	1.3%	top	0.5%	JouwBaan	0.9%	Directioners	0.5%
world	1.2%	topNews	0.5%	vacature	0.9%	JIMIN	0.5%
MAGA	0.8%	ISIS	0.4%	KCA	0.9%	PRODUCE101	0.5%
health	0.8%	Merkelmussbleiben	0.4%	Psychic	0.8%	VoteMainFPP	0.5%
local	0.7%	IslamKills	0.4%	RT	0.8%	Werk	0.4%
BlackLivesMatter	0.7%	breaking	0.4%	Libertad2016	0.6%	dts	0.4%

Account Evolution

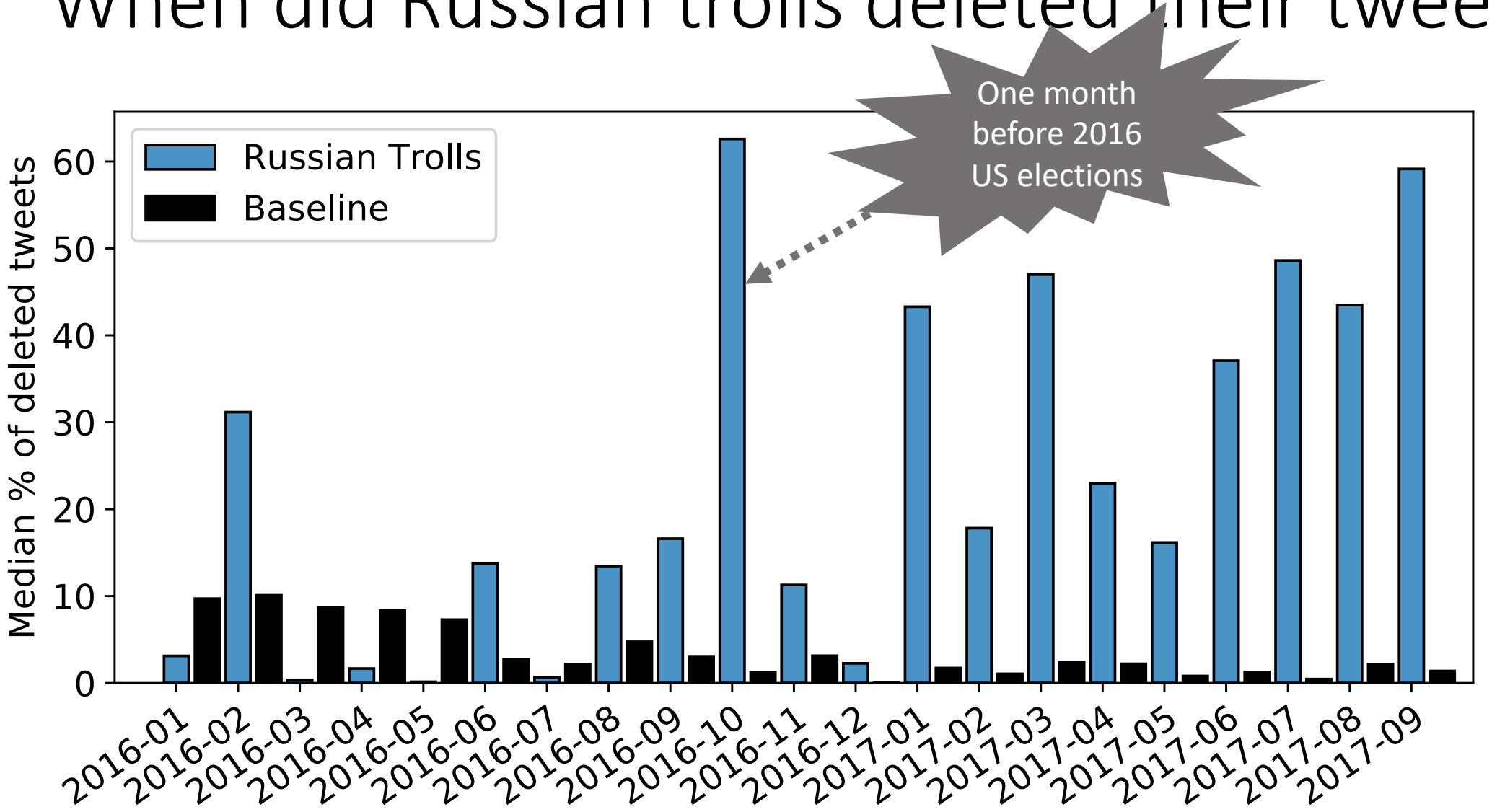
Do Russian trolls change their screen names?

- 9% of the Russian troll accounts changed their screen name
  - Up to 4 times per account
- E.g., from “OnlineHouston” to “HoustonTopNews”
  - Clear attempt to pose as local news outlet
- In our baseline dataset 19% of the accounts changed their screen name
  - Up to 11 times per account

Do Russian trolls delete their tweets?



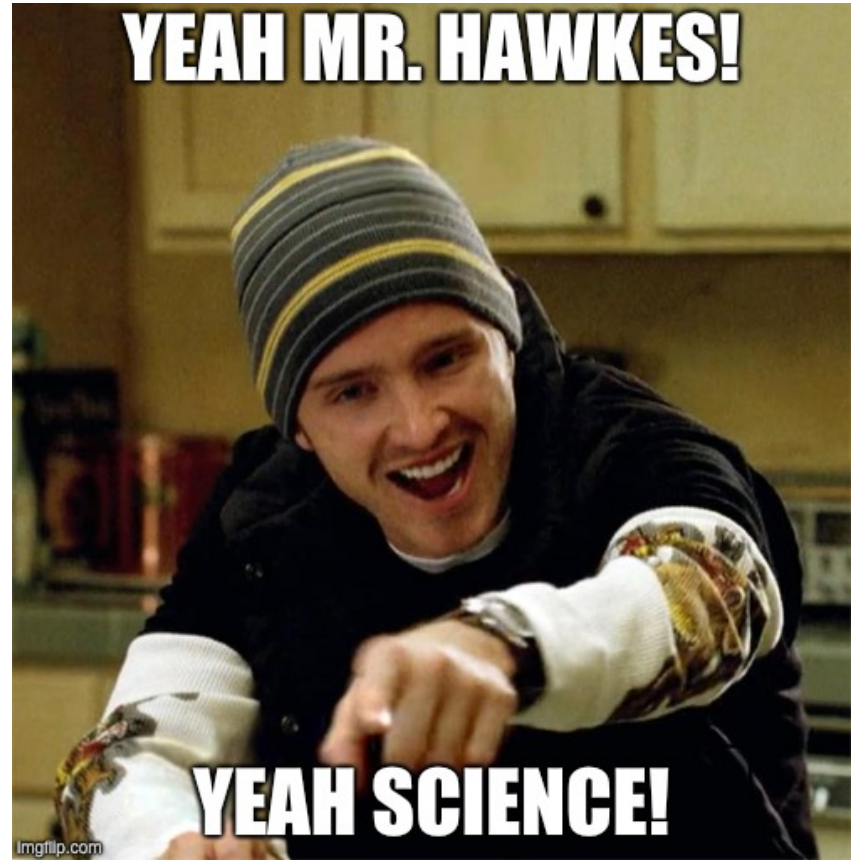
When did Russian trolls deleted their tweets?



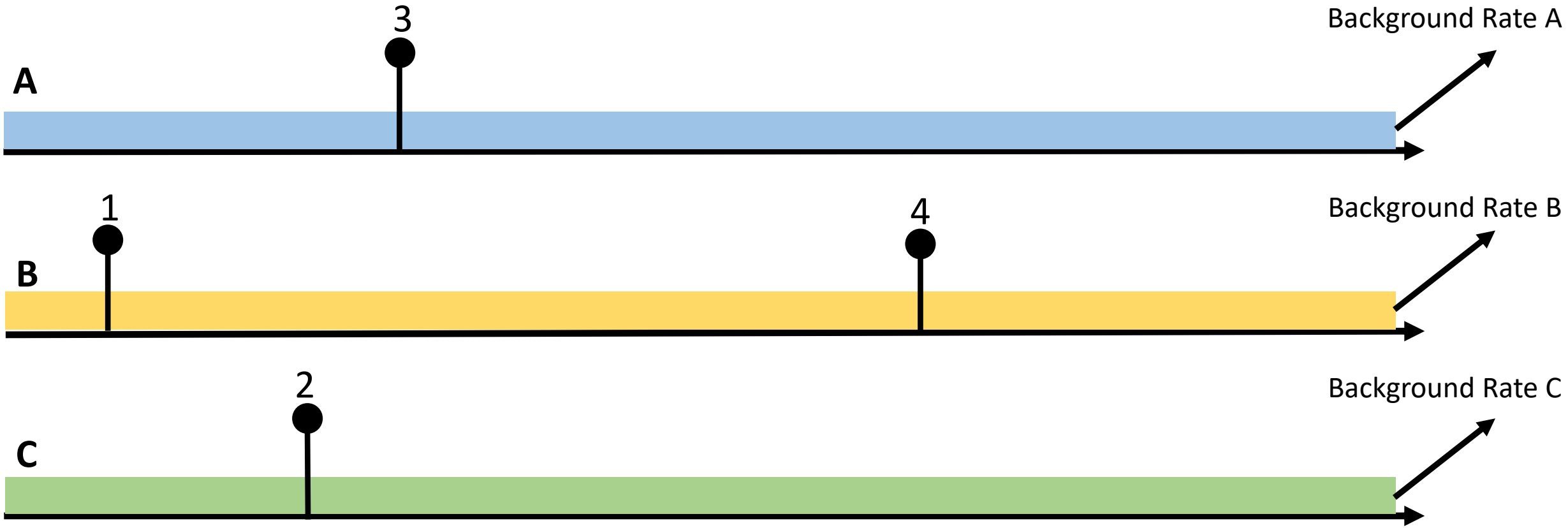
Influence Estimation

How to *quantify* the influence?

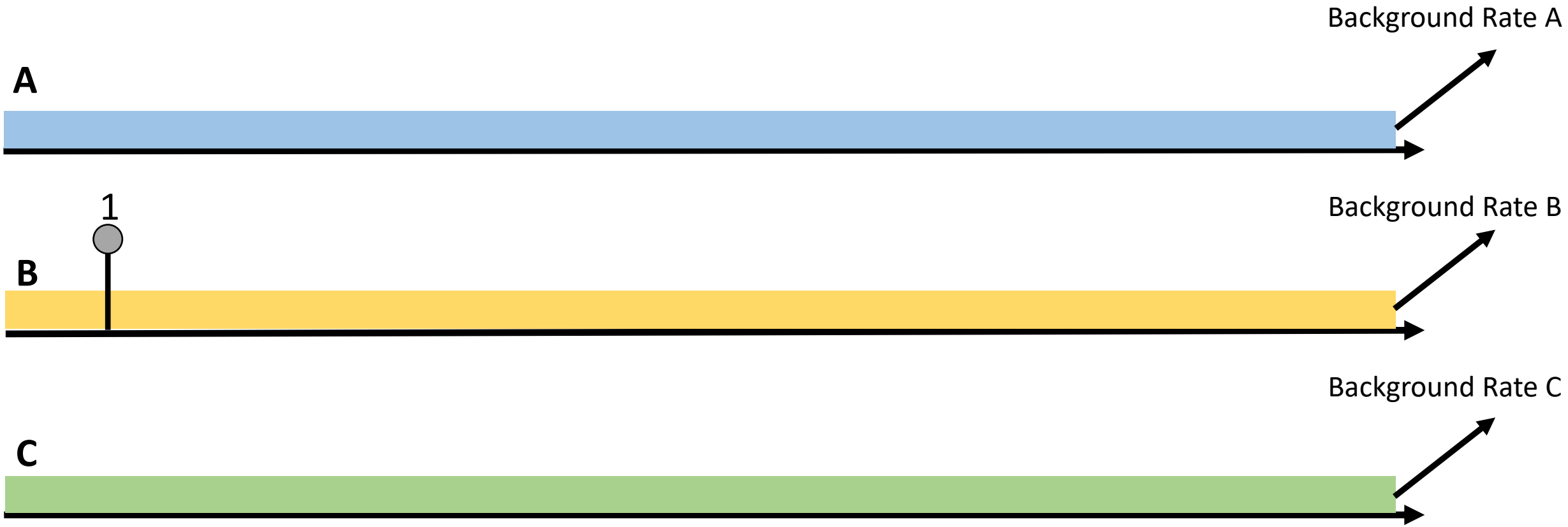
- *Hawkes processes*
- Assume K processes
  - Each with a rate of events (i.e., posting of a URL), called the *background rate*
- An event can cause *impulse responses* in other processes
  - Increases the rates of other processes for a period of time
- Enables us to assess root cause of events



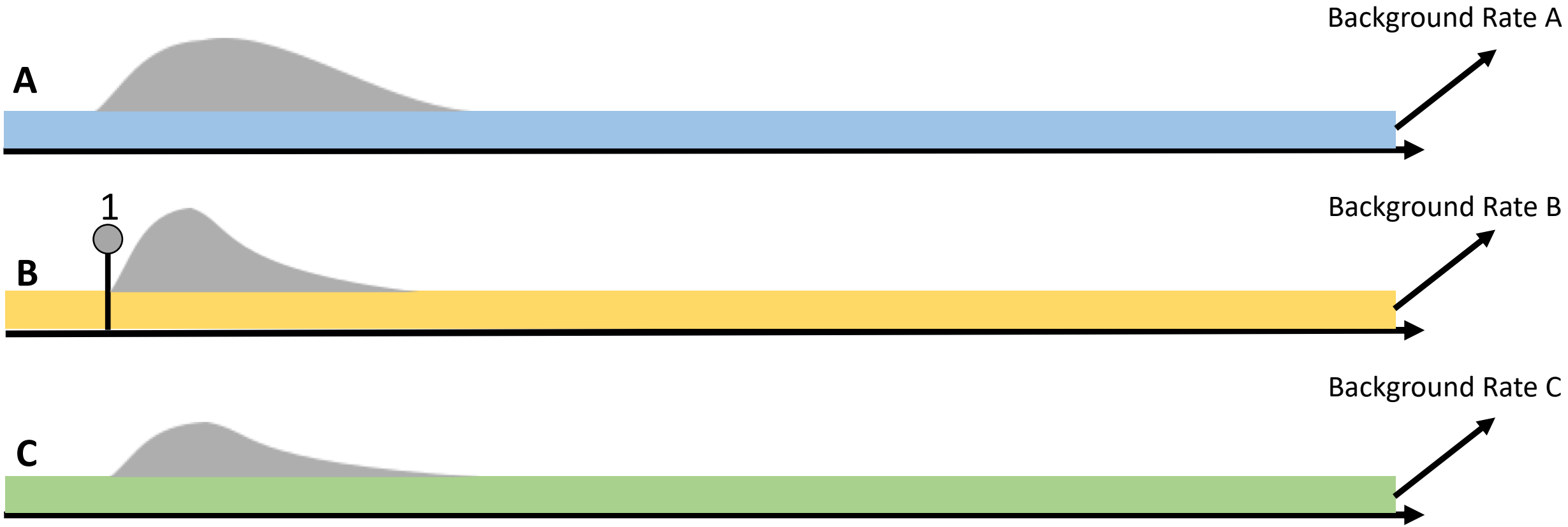
Hawkes processes example



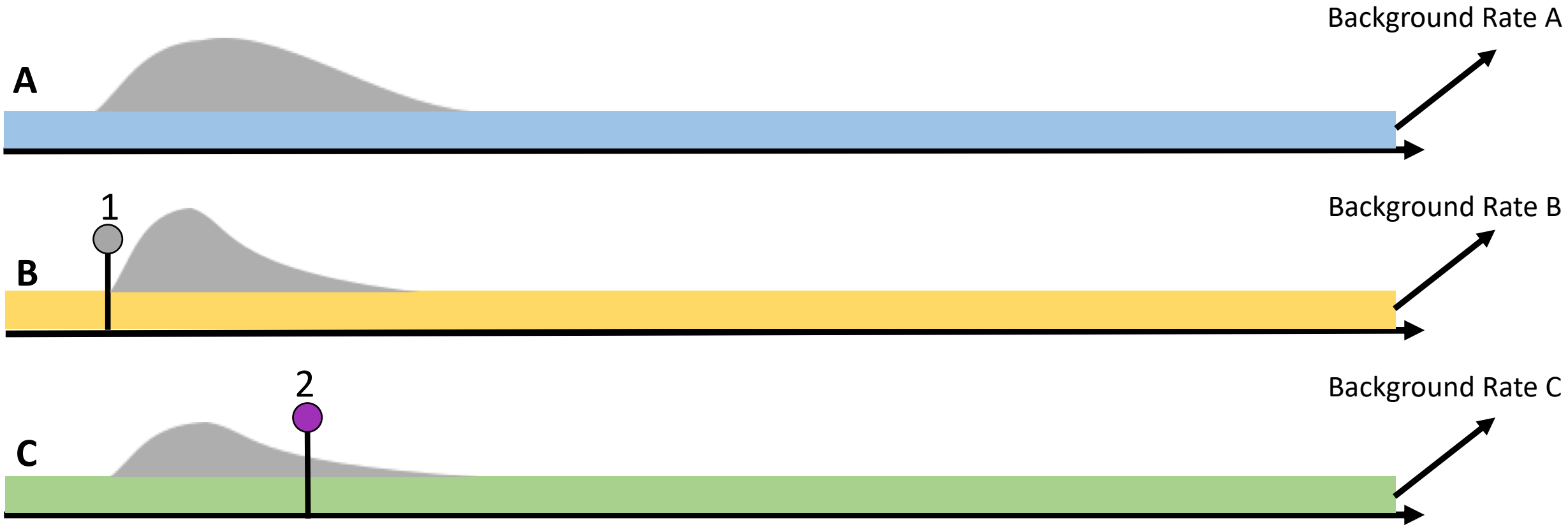
Hawkes processes example



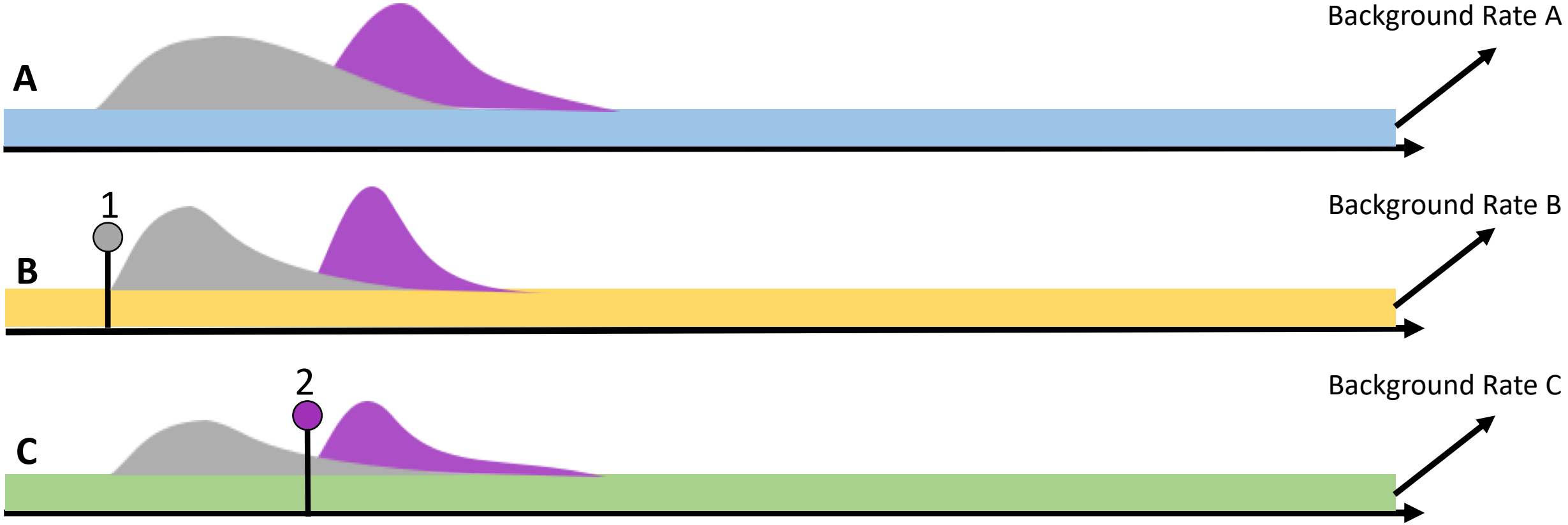
Hawkes processes example



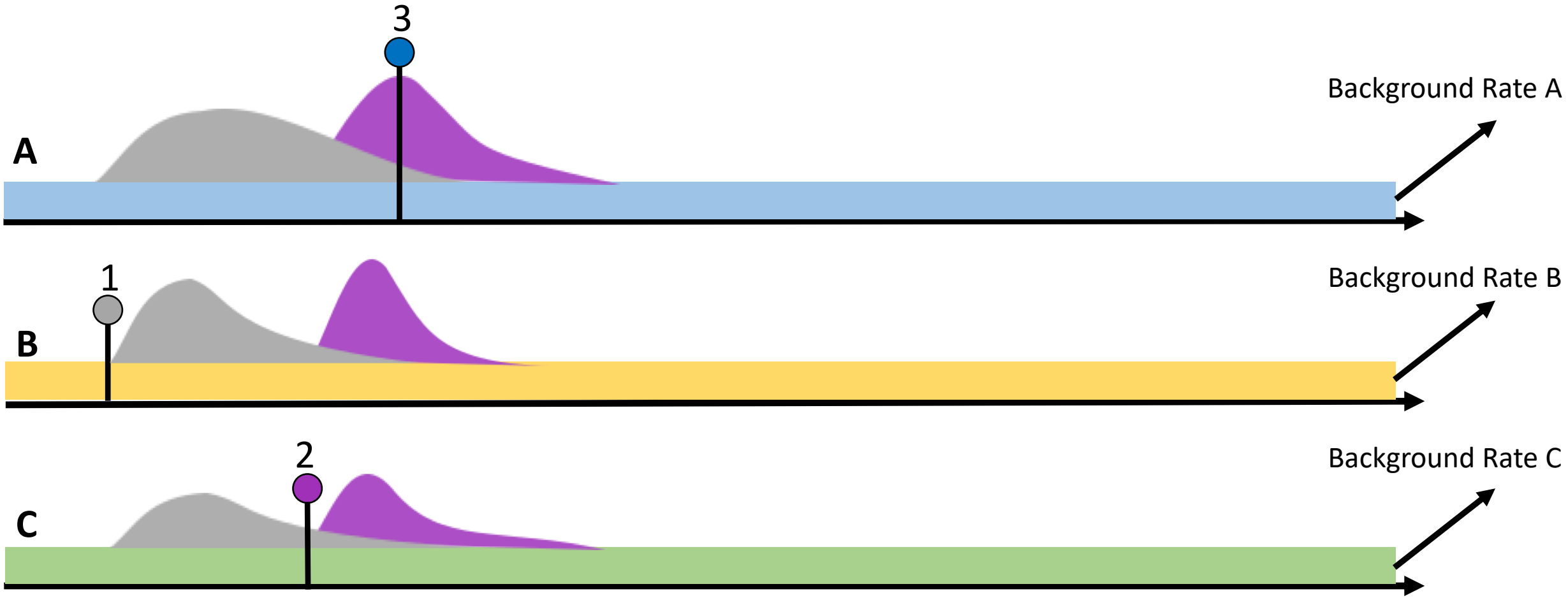
Hawkes processes example



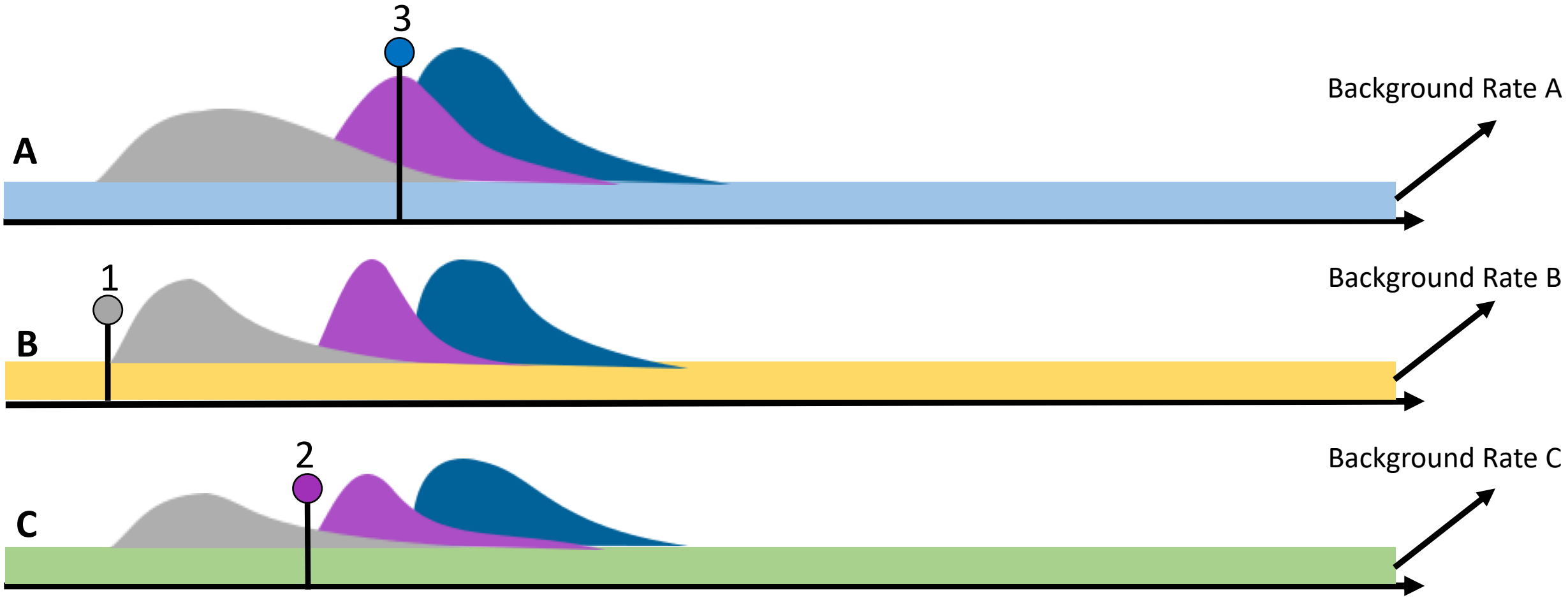
Hawkes processes example



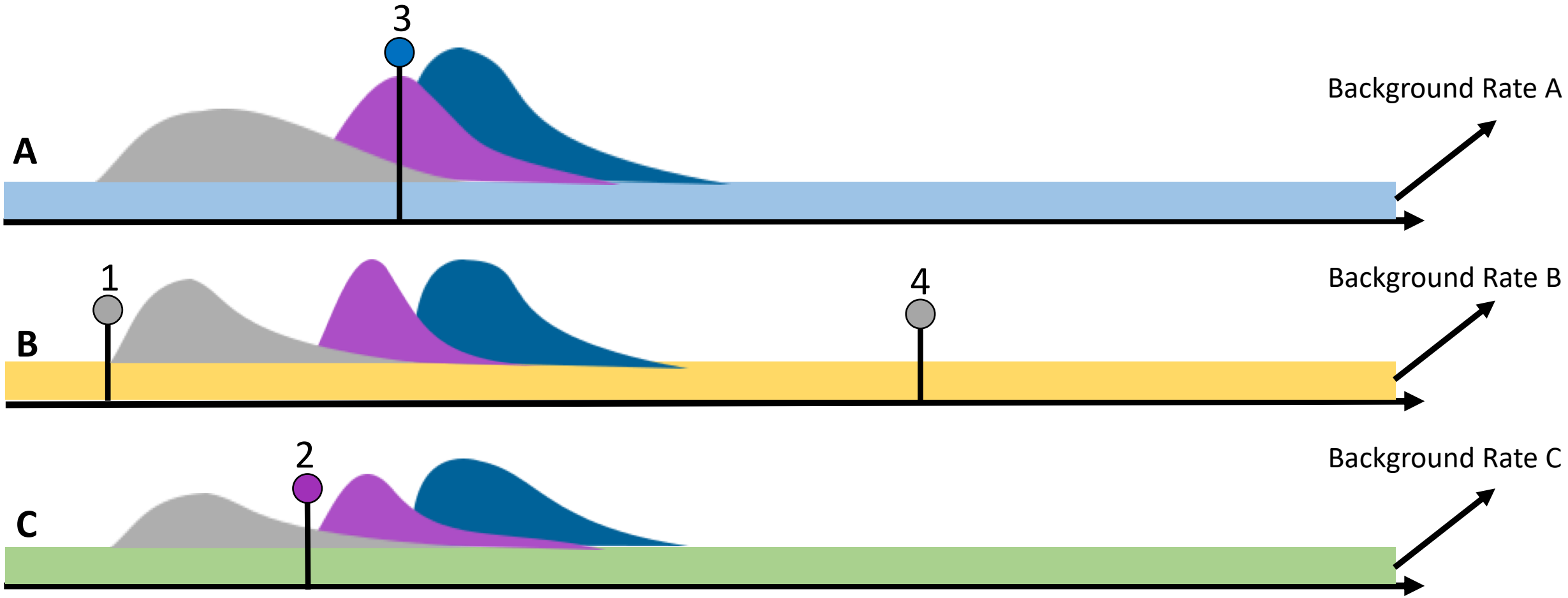
Hawkes processes example



Hawkes processes example



Hawkes processes example

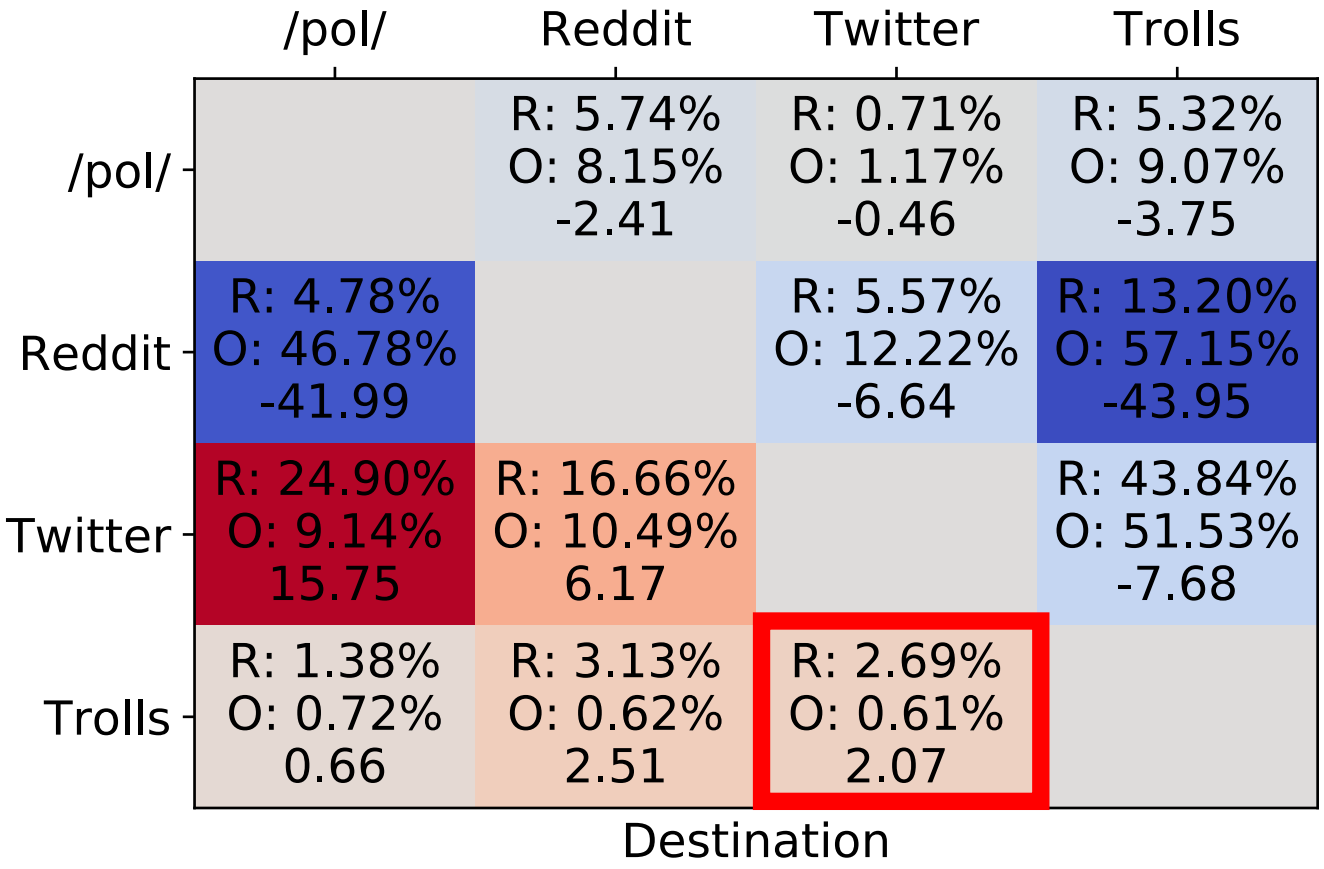


For our purposes

- Hawkes model with 4 processes
  - One for each platform/community of users (/pol/, Reddit, Twitter, Russian trolls)
- Use the a list of 99 news outlets from Zannettou et al. (IMC'17) to extract the URLs in each community.
- Distinct model for each URL; fit each model with Gibbs sampling
- Calculate the influence of each community

Influence results

- 1K trolls caused 2.6% of all Russian state-sponsored news outlets URLs (i.e., RT) on Twitter's 1%.
- 1K trolls caused 0.6% of all other news outlets URLs on Twitter's 1%.



Conclusions

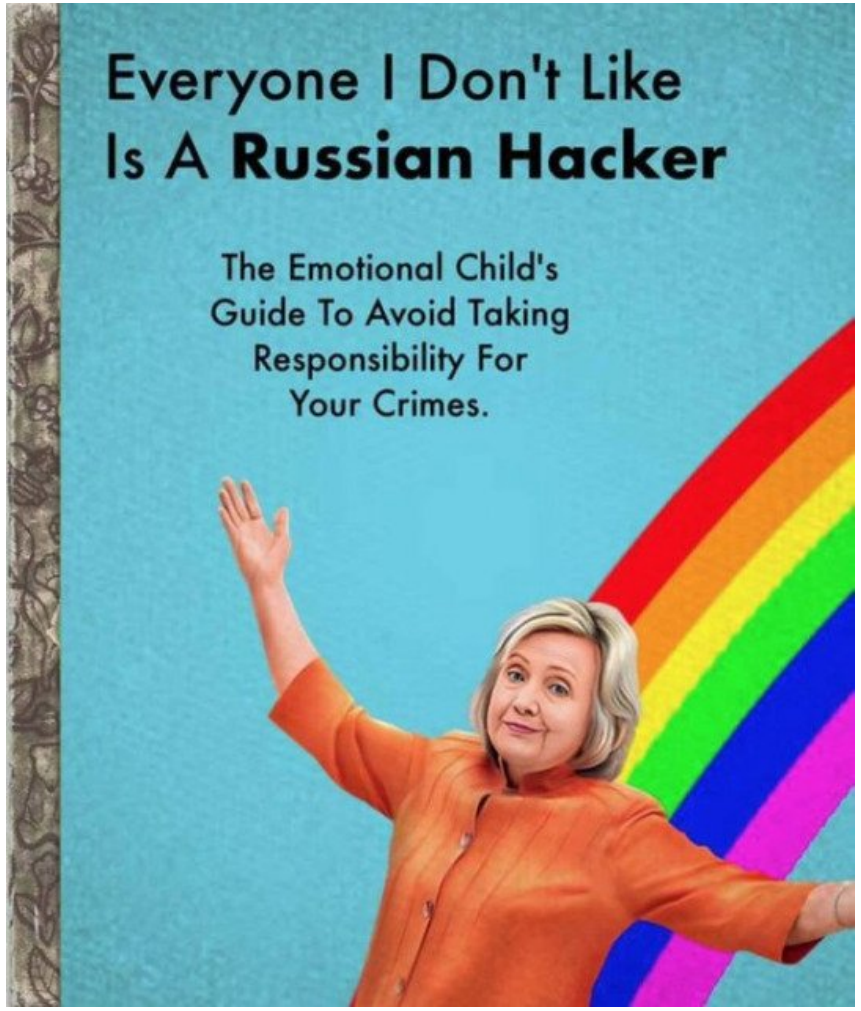
- We find differences in the use of the Twitter platform between trolls and random users
- Trolls seem to reset their “personas” by changing names and deleting tweets
- Particularly influential in spreading Russian state-sponsored URLs on Twitter and other platforms

Follow-up Related Work

- Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G. and Blackburn, J., 2018. Who let the trolls out? towards understanding state-sponsored trolls. *arXiv preprint arXiv:1811.03130* **(to appear at Websci'19).**

- Zannettou, S., Bradlyn, B., De Cristofaro, E., Stringhini, G. and Blackburn, J., 2019. Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter. *arXiv preprint arXiv:1901.05997* ***(under submission).***

Questions?


Everyone I Don't Like
Is A **Russian Hacker**
The Emotional Child's
Guide To Avoid Taking
Responsibility For
Your Crimes.


Days without winning
0 0 0 0


Raise your hand if you're under
investigation!