Hawkes Processes on Social Media and Mass Media - a Case Study of the #BlackLivesMatter Movement in the Summer of 2020

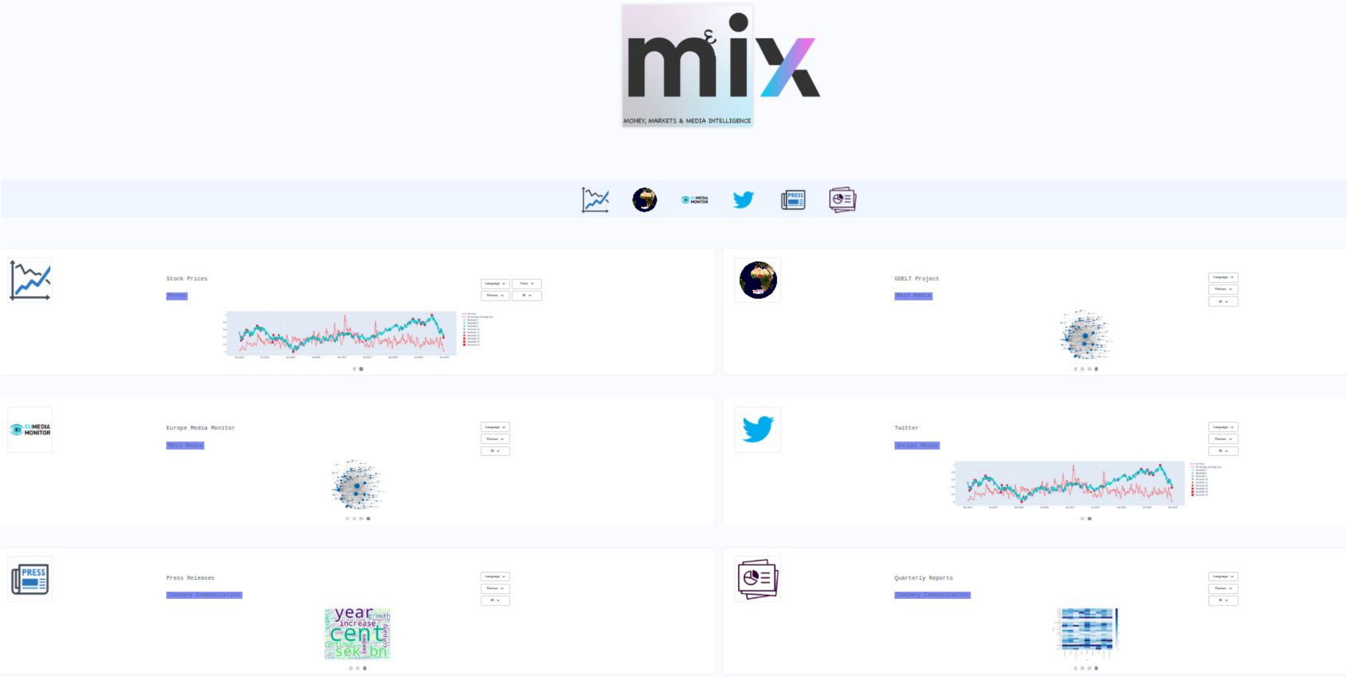
Alfred Lindström

Master thesis in applied mathematics and statistics, Uppsala University, 2021
Exjobb of Combient Mix AB

Supervisor: Raazesh Sainudiin

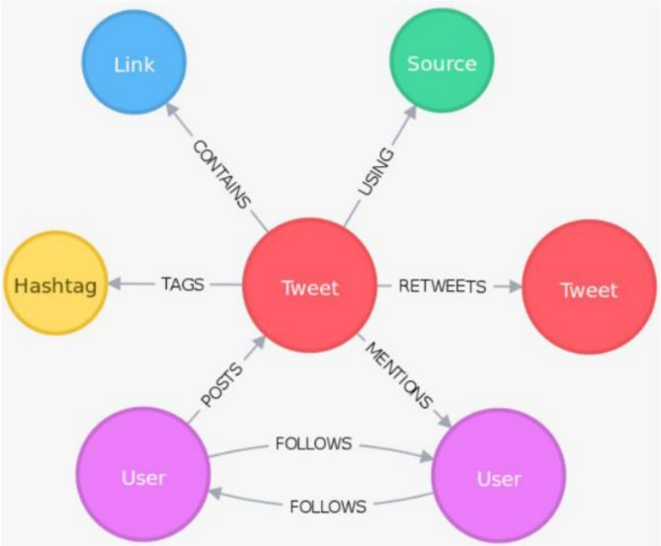
Context: “*money, markets & media intelligence*” is an ongoing project at Combient Mix AB

This is the media part of money, markets & media intelligence for Combient Mix AB



What is Twitter?

- Micro-blog where users share so called “tweets”
  - Short text messages
  - Media content such as videos and pictures
  - URLs
- User base consists of both public users such as politicians, journalists, and companies, and private users.
- Asymmetrical social media: Users may interact with each other without being friends.
- Users may tag their posts using hashtags, and also mention other users.



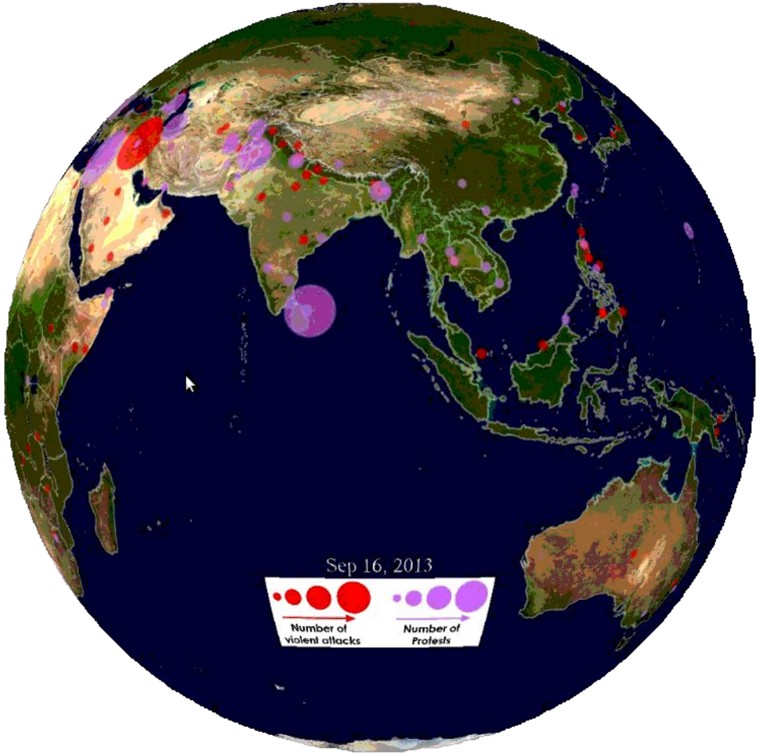
What is Twitter?

What is Twitter?

GDELT

"The GDELT Project is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day."

- Parses *records* e.g. news articles.
- Uses a coding framework that identifies events and actors being reported in these records from mass media.



GDELT - Coding

Example:

Sentence in a record:

“***President Reagan*** has ***threatened*** further action against ***the Soviet Union*** in an international television program beamed by satellite to more than 50 countries.”

GDELT - Coding

Example:

Sentence in a record:

“***President Reagan*** has ***threatened*** further action against ***the Soviet Union*** in an international television program beamed by satellite to more than 50 countries.”

- The act of *threatening* is identified as the event and given a code.

GDELT - Coding

Example:

Sentence in a record:

“***President Reagan*** has ***threatened*** further action against ***the Soviet Union*** in an international television program beamed by satellite to more than 50 countries.”

- The act of *threatening* is identified as the event and given a code.
- *President Reagan*, and *the Soviet Union* are identified as the relevant actors.

GDELT - Databases

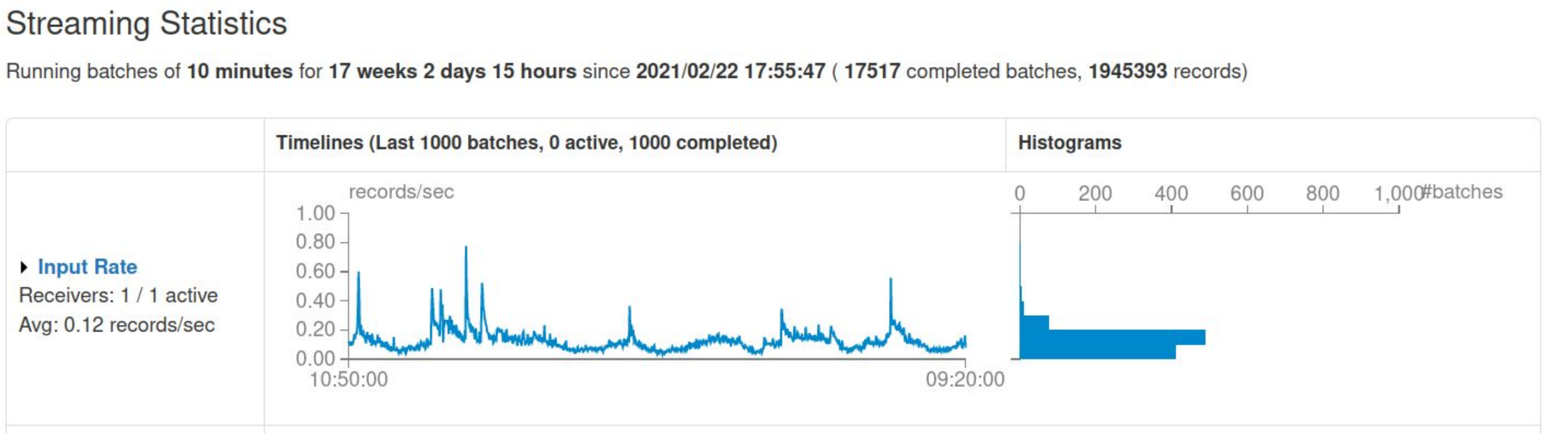
GDELT fundamentally consists of two datasets:

1) Global Knowledge Graph - GKG
    a) Contains the sources for the news being parsed.
    b) Updated every 15 minutes.
2) The Event Database
    a) Contains the events coded from GKG database.

Data Engineering for Data Science and Analytics - I

Infrastructure for Twitter data:

- A streaming job monitors all companies in the Combient network with active Twitter-accounts, along with their competitors Twitter accounts.
- The Tweets represented as raw .jsons are directly stored in a Delta Lake.
- Via schema inferring, data can be handled seamlessly in a Spark Context.



Data Engineering for Data Science and Analytics - II

- Twitter API (Application Programming Interface):
  - One tweet is represented as a .json-file
  - Main objects are **tweet-object** and **user-object**.

Data Engineering for Data Science and Analytics - II

- Twitter API (Application Programming Interface):
  - One tweet is represented as a .json-file
  - Main objects are **tweet-object** and **user-object**.

User object		
Attribute	Type	Description
id	Int64	The unique integer representation of the user.
screen_name	String	The screen name, also known as handle of the user.
followers_count	Int	The number of followers the user has.
friends_count	Int	The number of users the user follows.

Data Engineering for Data Science and Analytics - II

Tweet object		
Attribute	**Type**	**Description**
created_at	String	UTC-time when the tweet was created.
id	Int64	The unique integer representation of the tweet.
text	String	The textual content of the tweet.
in_reply_to_status_id	Int64	If the tweet is a reply to another tweet, the field will contain the tweet-ID of that tweet. Otherwise null.
in_reply_to_user_id	Int64	If the tweet is a reply to another tweet, the field will contain the user-ID of that tweet. Otherwise null.
user	User Object	All information of the user of the tweet.
quoted_status	Tweet Object	If the tweet is a quote tweet, all information of the original tweet will be contained in this field. Otherwise null
retweeted_status	Tweet Object	If the tweet is a retweet, all information of the original tweet will be contained in this field. Otherwise null

Data Engineering for Data Science and Analytics - II

Tweet object		
Attribute	Type	Description
created_at	String	UTC-time when the tweet was created.
id	Int64	The unique integer representation of the tweet.
text	String	The textual content of the tweet.
in_reply_to_status_id	Int64	If the tweet is a reply to another tweet, the field will contain the tweet-ID of that tweet. Otherwise null.
in_reply_to_user_id	Int64	If the tweet is a reply to another tweet, the field will contain the user-ID of that tweet. Otherwise null.
user	User Object	All information of the user of the tweet.
quoted_status	Tweet Object	If the tweet is a quote tweet, all information of the original tweet will be contained in this field. Otherwise null
retweeted_status	Tweet Object	If the tweet is a retweet, all information of the original tweet will be contained in this field. Otherwise null

Note:

- We get no information on the network structure between users, i.e., how users follow each other.
- Retweeted_status only points to the original tweet.

Case study

The Protest taking place after the killing of George Floyd last summer.

Why? Great example of the interrelationship of mobilization on social media, real world events and mass media.

Events

What happened?

- The death of George Floyd on the 25th of May 2020
  - The event was caught on camera by passers-by, and went viral on Facebook the same night.
- Largest protests in U.S. history
  - Mobilization under the hashtag #BlackLivesMatter
  - Protests also spread internationally

The #BlackLivesMatter-movement

What is Black Lives Matter?

- A decentralized grass-root movement that began on social media, using the hashtag #BlackLivesMatters.
- Founded in the wake of the shooting of Trayvon Martin, July 2013.
- Main issues is that of advocating against police brutality toward African-Americans, and policy issues related to racial injustices.
- Counter movements #AllLivesMatter, and #BlueLivesMatter has risen up as a response.

The #BlackLivesMatter-movement

What is Black Lives Matter?

- A decentralized grass-root movement that began on social media, using the hashtag #BlackLivesMatters.
- Founded in the wake of the shooting of Trayvon Martin, July 2013.
- Main issues is that of advocating against police brutality toward African-Americans, and policy issues related to racial injustices.
- Counter movements #AllLivesMatter, and #BlueLivesMatter has risen up as a response.

The decentralized nature of the BLM-movement, and the way social media has played a key part in its development, motivates our choice to analyze Twitter-data.

Case study

The Protest taking place after the killing of George Floyd last summer.

- 41.8 million collected tweets regarding the Black Lives Matter-movement, along with the smaller counter movements Blue Lives Matter (pro-police movement), and All Lives Matter.
- Tweets from the beginning of the movement in 2013 to the last of June 2020.

BLM Dataset

How was the data handled?

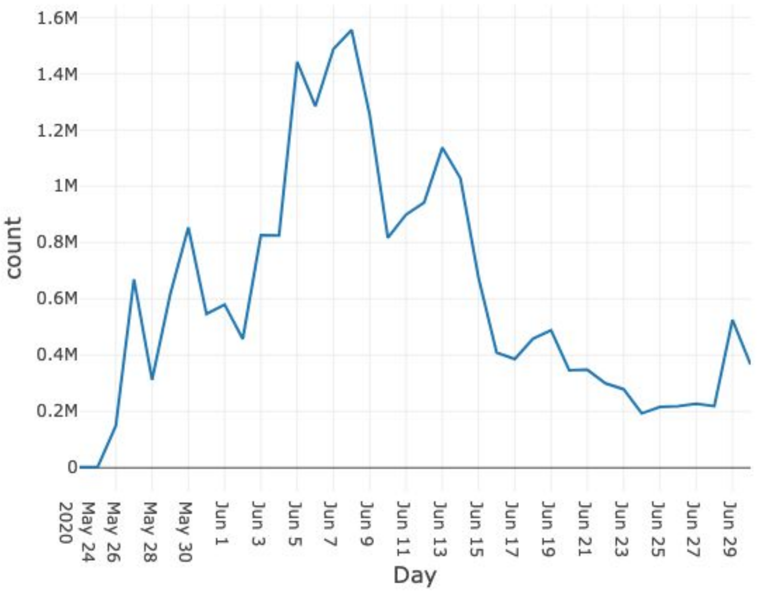
By Twitter Terms and Agreement, Tweets are not allowed to be stored and shared publicly.

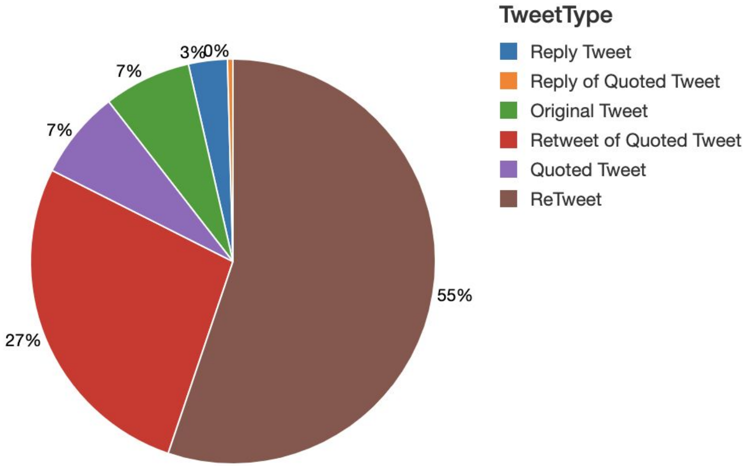
- To share data, one shares the relevant Tweet ID for each Tweet.
- From these IDs, one requests the Tweet data using Twitter credentials.
    - This was done using Python library `twarc`.
    - Different schema for the .json-files, so inferring had to be redone to able to handle the data in Databricks.
    - Resulted in a new infrastructure to get Twitter data retroactively into Databricks.

BLM Dataset

Twitter Data from May 24th - June 30th 2020

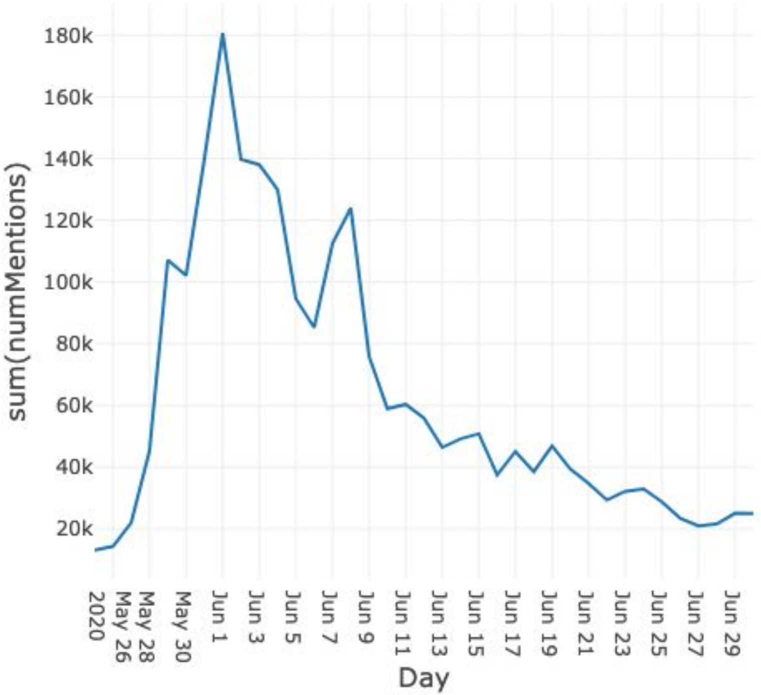
During this time period 23346745 tweets by 7111140 unique users were collected. 4101080 were original tweets





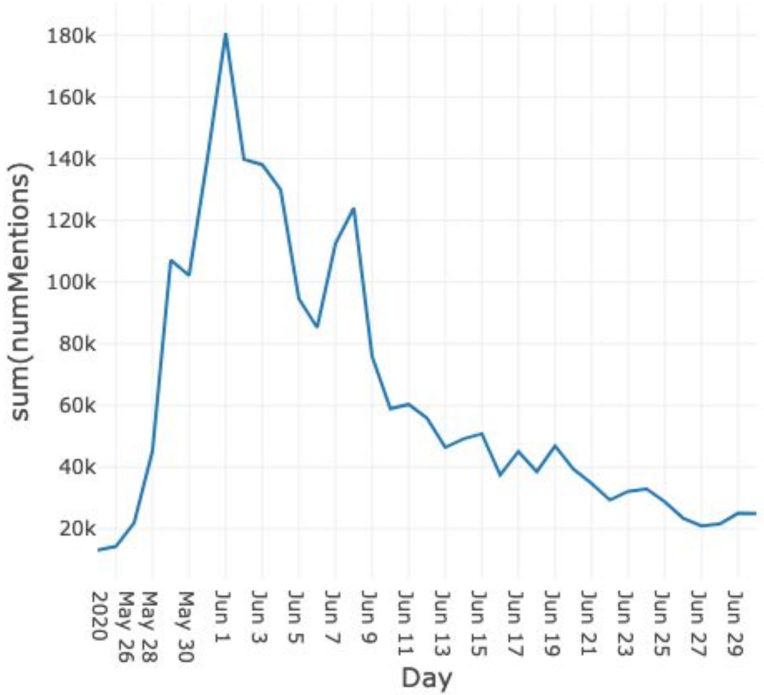
GDELT

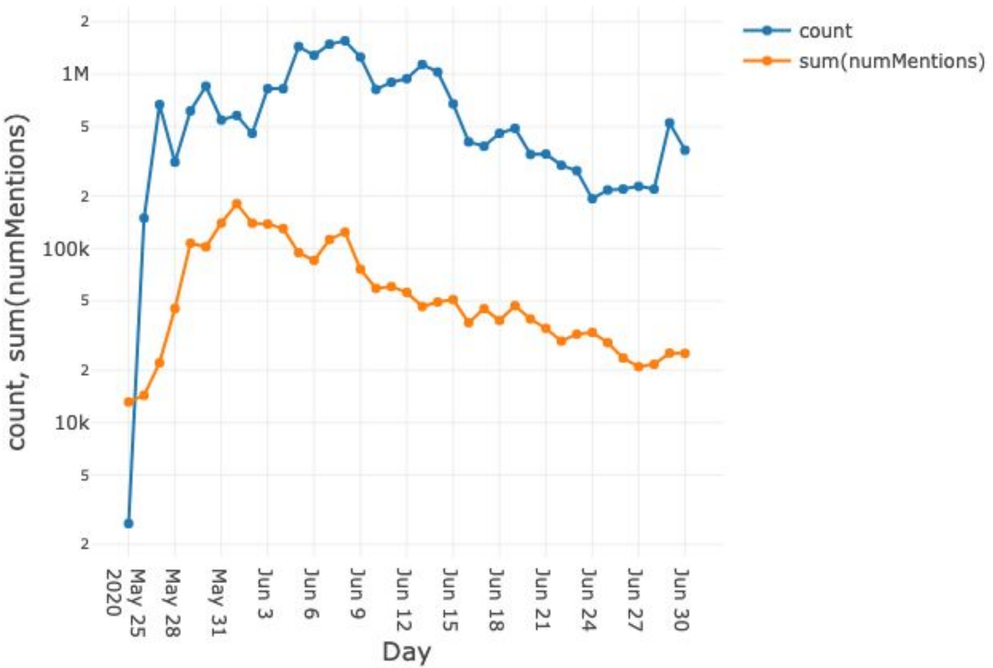
Records reporting protests during the same time period



GDELT

Records reporting protests during the same time period

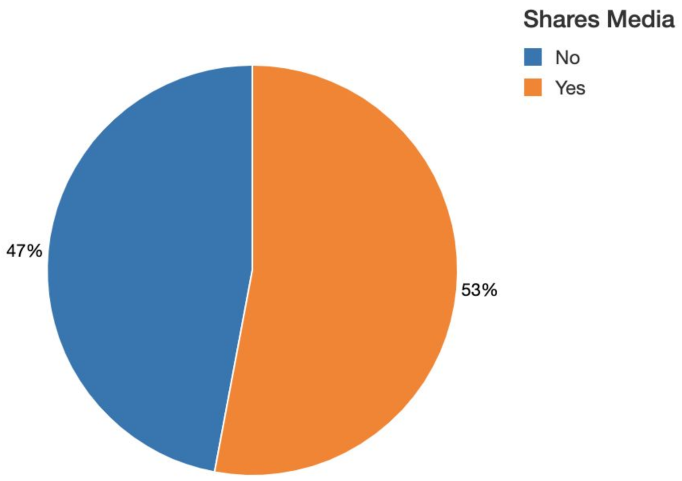




Role of Media-sharing

One initial idea was to link URLs to news articles from GDELT with shared URLs from Twitter. However, users shared more original media in favor of news articles.

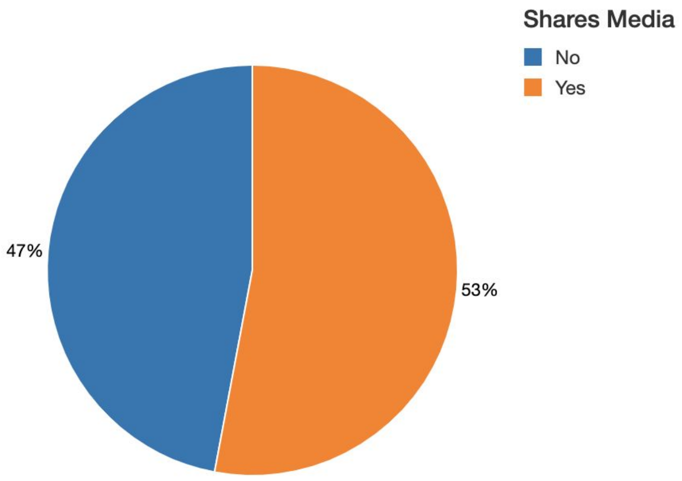
Original tweets with 1000 or more retweets:



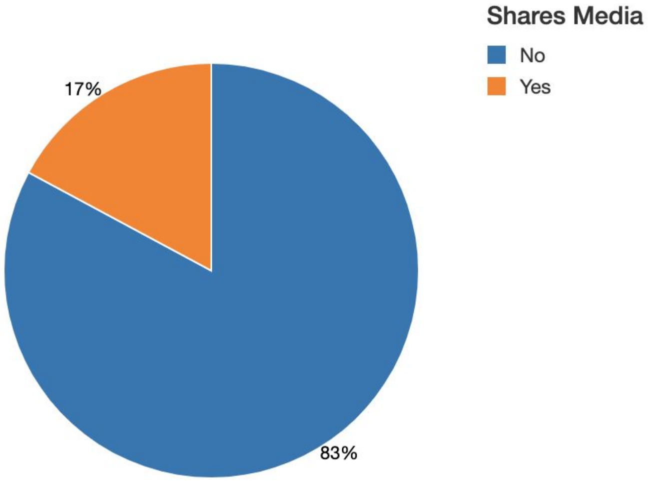
Role of Media-sharing

One initial idea was to link URLs to news articles from GDELT with shared URLs from Twitter. However, users shared more original media in favor of news articles.

Original tweets with 1000 or more retweets:



Total of all original tweets:



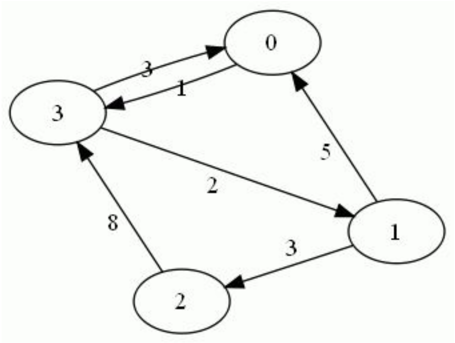
Retweet-network

How did users interact during this timeframe?

Retweet-network

How did users interact during this timeframe?

Let **G** be a weighted directed graph, where the vertices are users, and for the *n* times user *v* retweets user *u*, we add edge (u,v) with weight *n.*

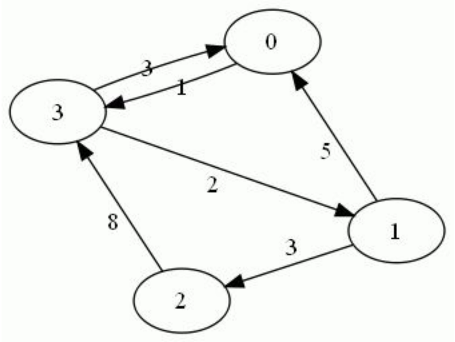


1. 85.6% of the users in **G** were in the same connected component.

Retweet-network

How did users interact during this timeframe?

Let **G** be a weighted directed graph, where the vertices are users, and for the *n* times user *v* retweets user *u*, we add edge (u,v) with weight *n.*

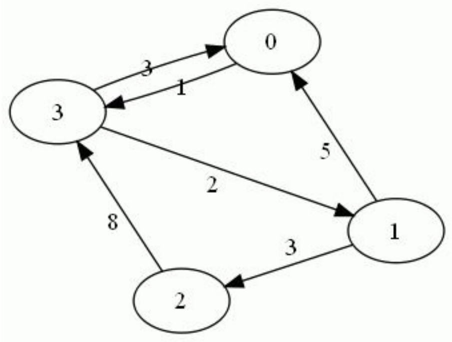


1. 85.6% of the users in **G** were in the same connected component.
2. From this the most retweeted users were identified.
   a. One pro-BLM journalist posting video content from the protests.
   b. One anti-BLM journalist posting video content from the protests.
   c. A few users with less than 1500 followers, but with over 100000 retweets.

Retweet-network

How did users interact during this timeframe?

Let **G** be a weighted directed graph, where the vertices are users, and for the *n* times user *v* retweets user *u*, we add edge (u,v) with weight *n.*



1. 85.6% of the users in **G** were in the same connected component.
2. From this the most retweeted users were identified.
   a. One pro-BLM journalist posting video content from the protests.
   b. One anti-BLM journalist posting video content from the protests.
   c. A few users with less than 1500 followers, but with over 100000 retweets.
3. A label propagation algorithm for community detection was run and two interesting communities were identified. One relating to BLM-tweets about the protests, and one relating to All/Blue Lives Matter.

Questions

I. Can we model the diffusion process of retweets being shared?

Questions

I. Can we model the diffusion process of retweets being shared?
II. What role does more influential users play in this process?

Questions

I. Can we model the diffusion process of retweets being shared?
II. What role does more influential users play in this process?
III. Can we say anything about the interaction between mass media and Twitter?

How to handle the given data?

- No social structure of the data was given.
- Both the GDELT and Twitter data is fundamentally points in time.

How to handle the given data?

- No social structure of the data was given.
- Both the GDELT and Twitter data is fundamentally points in time.

A natural choice would be to implement point processes.

Hawkes Processes - Self exciting point processes

Let $\mathcal{H}_t$ be the history of the events up to time *t* .

$$\lambda(t) = \mu + \sum_{t_i \in \mathcal{H}_t} \phi(t - t_i),$$

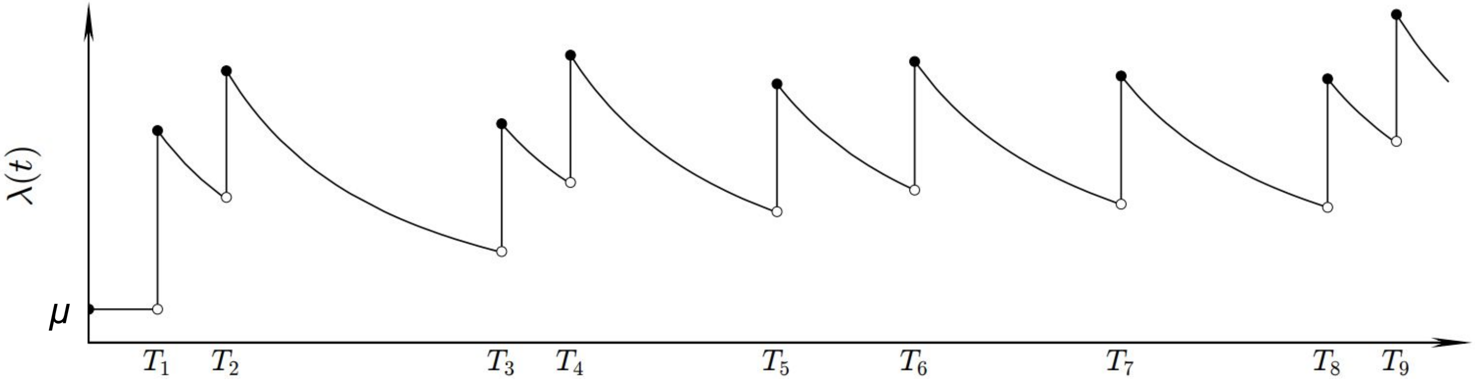
Where *μ* is the *baseline intensity*, and *ϕ(t)* is the *kernel*.

Hawkes Processes - Self exciting point processes

Let $\mathcal{H}_t$ be the history of the events up to time t.

$$\lambda(t) = \mu + \sum_{t_i \in \mathcal{H}_t} \phi(t - t_i),$$

Where μ is the *baseline intensity*, and $\phi(t)$ is the *kernel*.



Multivariate Hawkes Processes

Let *D* be the number of dimensions, and $\mathcal{H}^j_t$ the the history of events in dimension *j* up to time *t.* Then

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{t_{j,k} \in \mathcal{H}_t^j} \phi_{ij}(t - t_{j,k})$$

is the intensity for the *i*:th dimension.

Multivariate Hawkes Processes

Let *D* be the number of dimensions, and $\mathcal{H}^i_t$ the the history of events in dimension *i* up to time *t*. Then

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{t_{j,k} \in \mathcal{H}_t^j} \phi_{ij}(t - t_{j,k})$$

is the intensity for the *i*:th dimension.

- $t_{j,k}$ are timestamps in the *j*:th dimension

Multivariate Hawkes Processes

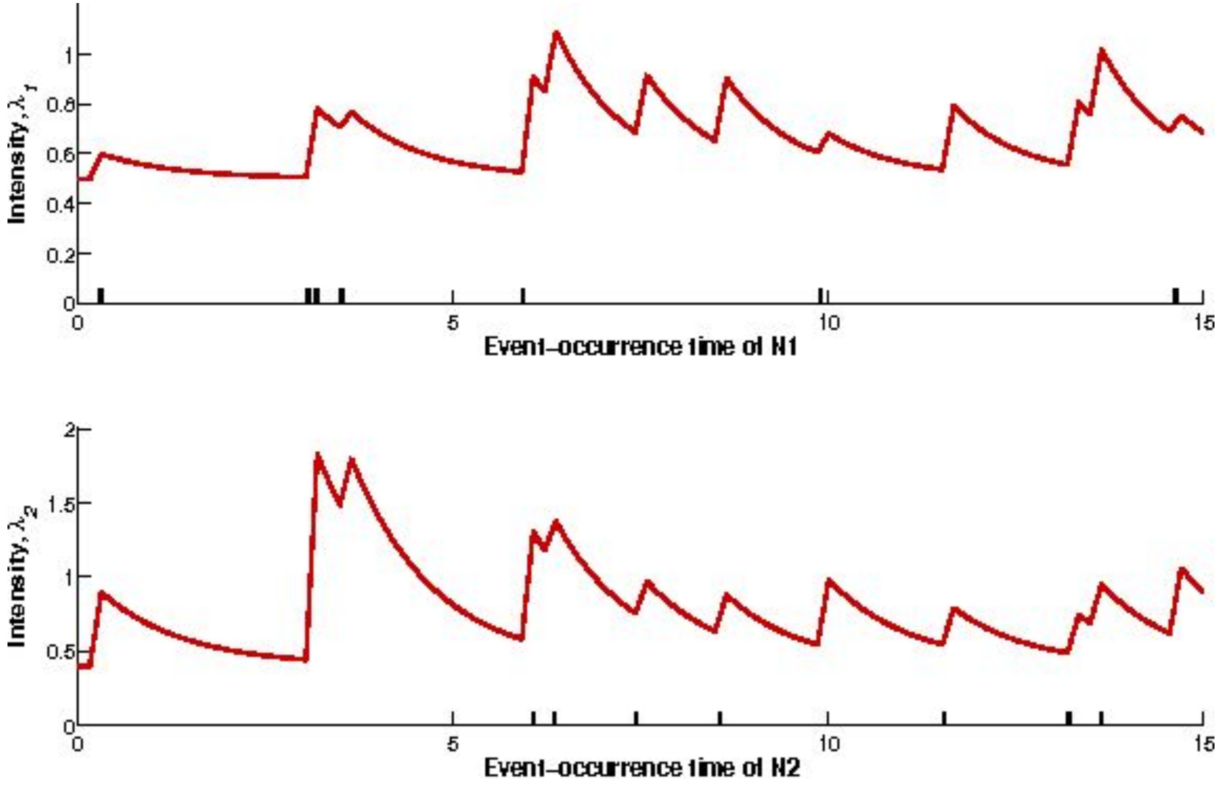
Let *D* be the number of dimensions, and $\mathcal{H}^i_t$ the the history of events in dimension *i* up to time *t.* Then

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{t_{j,k} \in \mathcal{H}_t^j} \phi_{ij}(t - t_{j,k})$$

is the intensity for the *i*:th dimension.

- $t_{j,k}$ are timestamps in the *j*:th dimension
- $\phi_{i,j}$ are the kernels - regulates how events in dimension *j* affects the intensity in dimension *i*

Multivariate Hawkes Processes



Interplay between mass media and Twitter

Notes on the exponential kernel:

$$\phi_{ij} = \alpha^{ij}\beta^{ij}\exp(-\beta^{ij}(t - t_k^j))$$

- $\alpha^{ij} > 0$ regulates how much of an impact of the intensity rate in dimension *i* an event in dimension *j* has.
- □$^{ij} > 0$ is the decaying parameter.
- The kernel is monotonically decreasing - thus events in the past will only affect the intensity when they are close in time.

Interplay between mass media and Twitter

The first 30 hours of when the protests were studied.

- Original Tweets during this time period was taken from the BLM-dataset
- Records mentioning protests were queried from GDELT.

Interplay between mass media and Twitter

The first 30 hours of when the protests were studied.

- Original Tweets during this time period was taken from the BLM-dataset
- Records mentioning protests were selected from GDELT.

For this model, a multivariate Hawkes process in two dimension with an exponential kernel was implemented:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{2} \sum_{t_{j,k} \in \mathcal{H}_t^j} \alpha^{ij} \beta^{ij} \exp(-\beta^{ij}(t - t_k^j))$$

The parameters were estimated using `Python`-library `ticks`, by fixing all decay parameters □ ^{ij} and then using the method of least-squares.

Interplay between mass media and Twitter

Notes on the exponential kernel:

$$\phi_{ij} = \alpha^{ij}\beta^{ij}\exp(-\beta^{ij}(t - t_k^j))$$

- α^{ij} *>0* regulates how much of an impact of the intensity rate in dimension *i* an event in dimension *j* has.
- □ ^{ij} *>0* is the decaying parameter.
- The kernel is monotonically decreasing - thus events in the past will only affect the intensity when they are close in time.

The parameters were estimated using `Python`-library `ticks`, by fixing all decay parameters □ ^{ij} and then using the method of least-squares.

Interplay between mass media and Twitter

H_0: There is no interplay (mutual excitation) between reported protests in media and Tweets regarding the BLM-movement, i.e., $\alpha_{12}=\alpha_{21}=0$

1. The assumption that all □ ^{ij} were equal was made.
    a. Different values for □ was then given for the fitting. The □ with the highest likelihood was chosen.
2. With a set □, bootstrapping by sampling from the GDELT-records and Tweets 100 times was done.

Interplay between mass media and Twitter

H_0: There is no interplay (mutual excitation) between reported protests in media and Tweets regarding the BLM-movement, i.e., $\alpha_{12}=\alpha_{21}=0$

- We get the 95% confidence intervals (0.013379, 0.030991) for α_{12} , (0.01131 ,0.022001) for α_{21}.
- Thus we do not reject H_0 according to the Wald test.

Note that for the means, $\alpha_{21} > \alpha_{12}$. This suggests that mass media has a larger effect on Twitter in this model.

	likelihood	baseline1	baseline2	a11	a12	a21	a22
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	178.581174	0.991167	1.004230	0.992748	0.024603	0.016649	0.969400
std	0.439215	0.000337	0.000199	0.002379	0.003866	0.003016	0.002535
min	177.561619	0.990209	1.003567	0.985194	0.013379	0.007766	0.963584
2.5%	177.766396	0.990478	1.003825	0.988841	0.017154	0.011311	0.965202
50%	178.631151	0.991146	1.004242	0.992610	0.024810	0.016794	0.969431
97.5%	179.440629	0.991800	1.004512	0.997414	0.030991	0.022001	0.974012
max	179.622568	0.992144	1.004758	0.999777	0.037128	0.023900	0.977783

Interplay between mass media and Twitter

This model is quite simple and should be interpreted as a first step.

Next steps:

1. Make a more formalized and well-defined problem statement.
2. Look at the Granger Causality to make better assumptions on predictive causality.

Modelling Retweet Cascades

I. Can we model the diffusion process of retweets being shared?
II. What role does more influential users play in this process?

Modelling Retweet Cascades

Retweet cascade - one original tweet along with all its retweet.

Modelling Retweet Cascades

Retweet cascade - one original tweet along with all its retweet.

Phenomena to capture:

1. **Word-to-mouth spread**: When a user shares a tweet, the tweet will organically find its way into new a set of new users, and so on.

Modelling Retweet Cascades

Retweet cascade - one original tweet along with all its retweet.

Phenomena to capture:

1. **Word-to-mouth spread**: When a user shares a tweet, the tweet will organically find its way into new a set of new users, and so on.
2. **The magnitude of influence**: Users with more followers tend to get more retweets.

Modelling Retweet Cascades

Retweet cascade - one original tweet along with all its retweet.

Phenomena to capture:

1. **Word-to-mouth spread**: When a user shares a tweet, the tweet will organically find its way into new a set of new users, and so on.
2. **The magnitude of influence**: Users with more followers tend to get more retweets.
3. **Memory over time**: Most of the retweeting by users happen when the users first see it in their timeline.

Modelling Retweet Cascades

Retweet cascade - one original tweet along with all its retweet.

Phenomena to capture:

1. **Word-to-mouth spread**: When a user shares a tweet, the tweet will organically find its way into new a set of new users, and so on.
2. **The magnitude of influence**: Users with more followers tend to get more retweets.
3. **Memory over time**: Most of the retweeting by users happen when the users first see it in their timeline.
4. **Content quality.**

Hawkes Processes for retweet cascades

$$\lambda(t) = \mu + \sum_{t_i \in \mathcal{H}_t} \phi(t - t_i),$$

Marked Power Law-kernel:

$$\phi^p(m_i, t) = \kappa m_i^\beta (t + c)^{-(1+\theta)}$$

Where each event along with a timestamp t_i also has a *mark* m_i which we let be number of followers the retweeting user has.

Hawkes Processes for retweet cascades - Kernels

$$\phi^p(m_i, t) = \kappa m_i^\beta (t + c)^{-(1+\theta)}$$

- *κ>0,* is interpreted as the quality of the tweet.

Hawkes Processes for retweet cascades - Kernels

$$\phi^p(m_i, t) = \kappa m_i^\beta (t + c)^{-(1+\theta)}$$

- *κ>0,* is interpreted as the quality of the tweet.
- *β>0,* determines how much of an impact the number of followers a user has influences the rate.

Hawkes Processes for retweet cascades - Kernels

$$\phi^p(m_i, t) = \kappa m_i^\beta (t + c)^{-(1+\theta)}$$

- *κ>0,* is interpreted as the quality of the tweet.
- *β>0,* determines how much of an impact the number of followers a user has influences the rate.
- *$(t+c)^{-(1+\theta)}$* , *θ,c >0* is monotonically decreasing, so that the relevancy of retweet dies out over time.

Hawkes Processes for retweet cascades - Kernels

$$\phi^p(m_i, t) = \kappa m_i^\beta (t + c)^{-(1+\theta)}$$

- *κ>0,* is interpreted as the quality of the tweet.
- *β>0,* determines how much of an impact the number of followers a user has influences the rate.
- $(t+c)^{-(1+\theta)}$, *θ,c >0* is monotonically decreasing, so that the relevancy of retweet dies out over time.

One retweet cascade is then modelled as a point process with intensity

$$\lambda(t) = \sum_{t_i \in \mathcal{H}_t} \alpha m_i^\beta (t + c)^{-(t+\theta)}$$

Hawkes Processes for retweet cascades - Estimating

Log-likelihood:

$$
\begin{aligned}
\mathcal{L}(\kappa, \beta, c, \theta \mid \mathcal{H}_{t_n}) &= \log P(\{(m_i, t_i), i = 1, ..., n\}) \\
&= \sum_{i=1}^{n} \log(\lambda(t_i)) - \int_0^T (\tau) \mathrm{d}\tau \\
&= \sum_{i=2}^{n} \log \kappa + \sum_{i=2}^{n} \log \left(\sum_{t_j < t_i} \frac{m_j^\beta}{(t_i - t_j + c)^{1+\theta}} \right) \\
&- \kappa \sum_{i=1}^{n} m_i^\beta \left[\frac{1}{\theta c^\theta} - \frac{(T + c - t_i)^{-\theta}}{\theta} \right]
\end{aligned}
$$

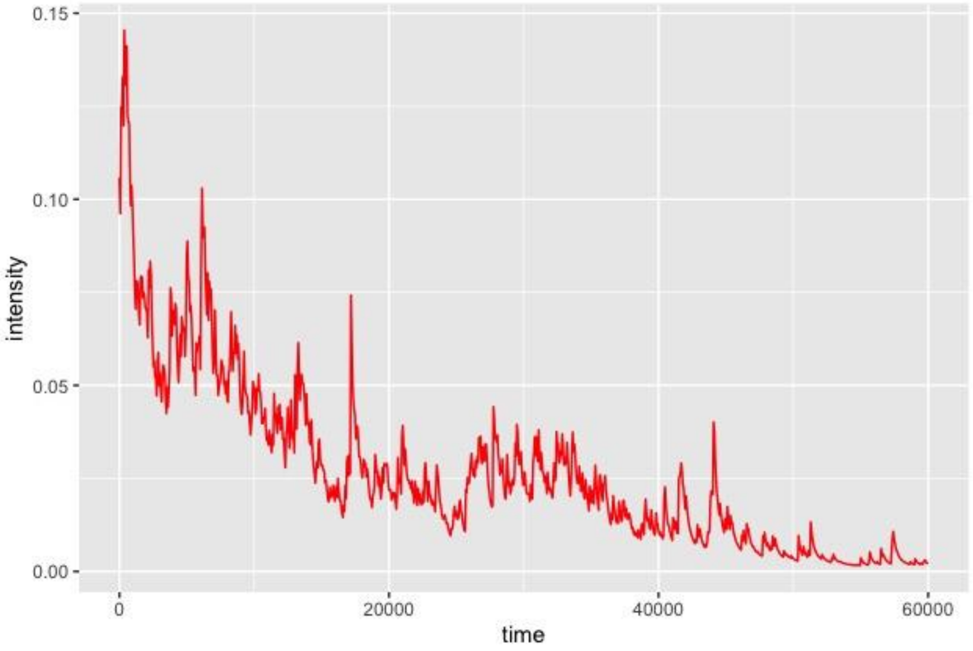
Non-linear, and optimized numerically using R-package `Evently`.

Hawkes Processes for retweet cascades - Estimating

Due to computational limitations, only retweet cascades at around 3500 tweets were able to be fitted.

Hawkes Processes for retweet cascades

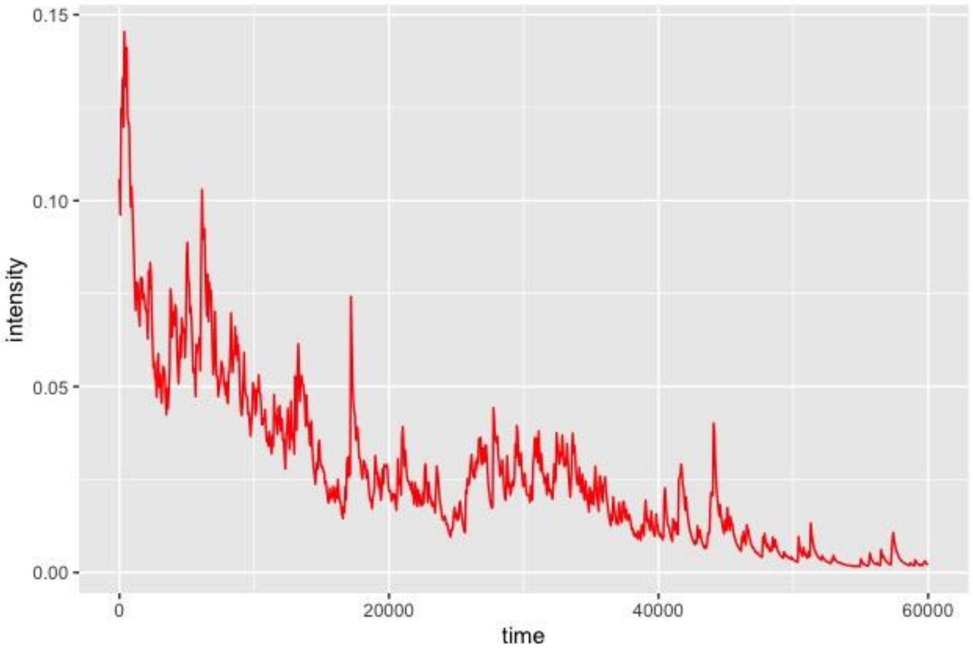
Cascades by prominent users from both BLM-movement and anti-BLM-movement were fitted.



Hawkes Processes for retweet cascades

Cascades by prominent users from both BLM-movement and anti-BLM-movement were fitted.

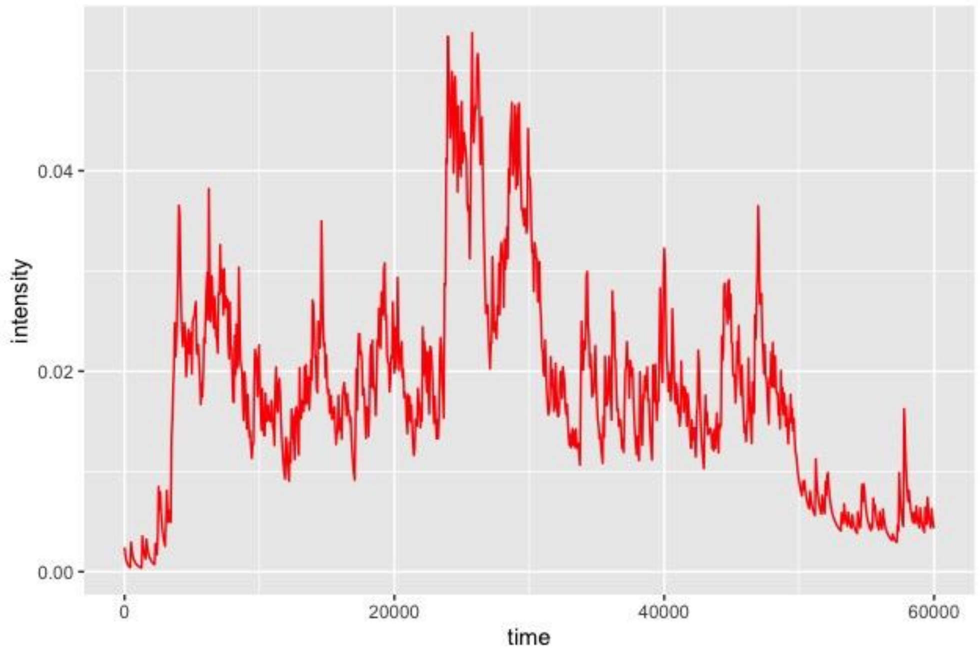
- Independent of political sympathies, these cascades looked quite similar.



Hawkes Processes for retweet cascades

More interesting were relatively large retweet cascades initiated by users with small followings.

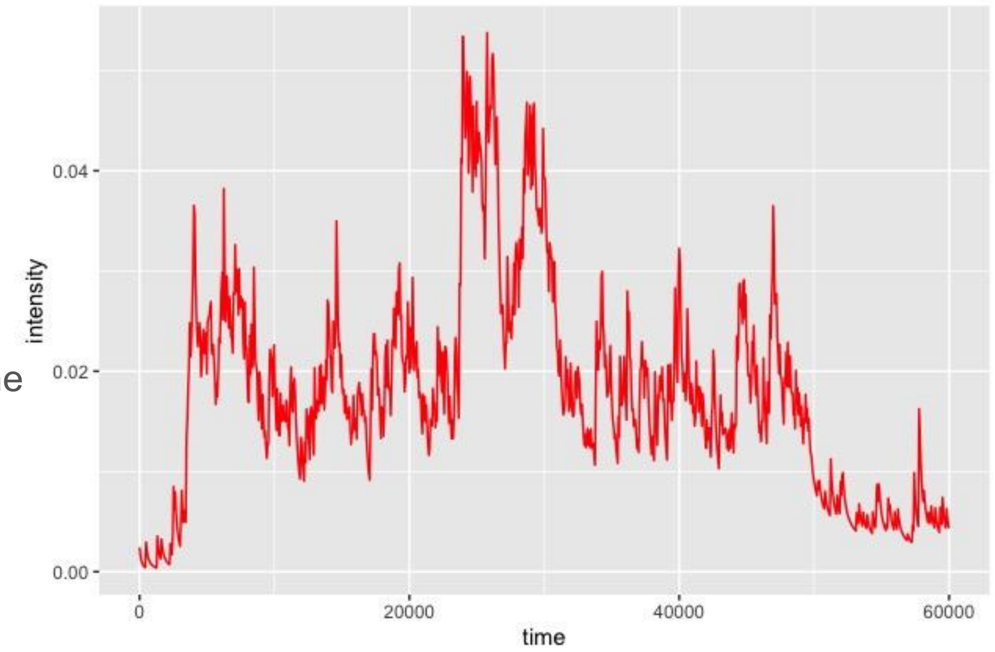
- Cascade with final size of 1500 retweets, initiated by a user with 268 followers.



Hawkes Processes for retweet cascades

More interesting were relatively large retweet cascades initiated by users with small followings.

- Cascade with final size of 1500 retweets, initiated by a user with 268 followers.

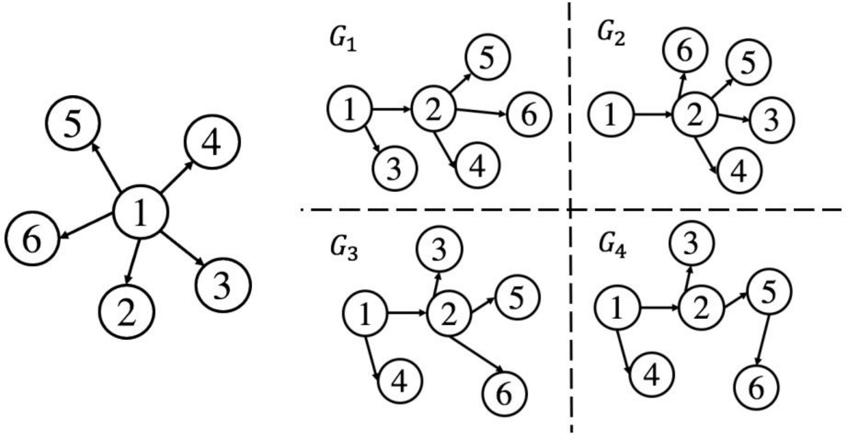


Cascades of this nature leads us into the next question: Impact of influential users.

Hawkes Processes for retweet cascades - Diffusion

Reminder: Due to Twitter API, we do not have access to the actual branching structure of a retweet cascade.

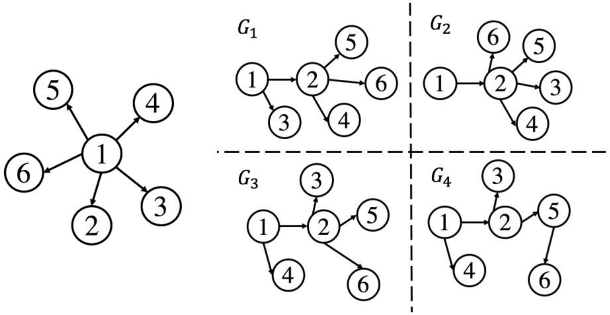
From the our model, an estimate of the probability of a tweet being a direct retweet of an earlier tweet (in the cascade) can be made.



Hawkes Processes for retweet cascades - Influence

Estimate of the probability that a tweet v_j is a direct retweet of v_i:

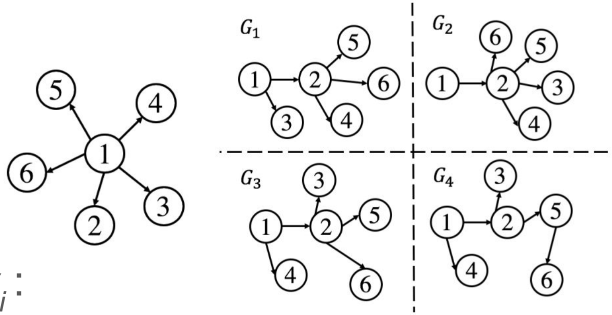
$$p_{ij} = \frac{\phi(m_i, t_j - t_i)}{\sum_{k=1}^{j-1} \phi(m_k, t_j - t_k)}$$



Hawkes Processes for retweet cascades - Influence

Estimate of the probability that a tweet v_j is a direct retweet of v_i:

$$p_{ij} = \frac{\phi(m_i, t_j - t_i)}{\sum_{k=1}^{j-1} \phi(m_k, t_j - t_k)}$$



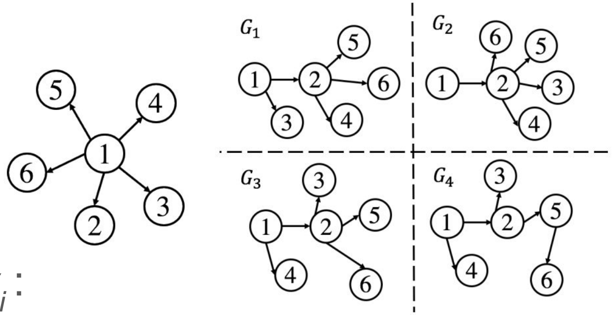
Pairwise influence, i.e. probability that v_j indirectly influences v_i:

$$r_{ij} = \sum_{k}^{j-1} r_{ik} p_{kj}$$

Hawkes Processes for retweet cascades - Influence

Estimate of the probability that a tweet v_j is a direct retweet of v_i:

$$p_{ij} = \frac{\phi(m_i, t_j - t_i)}{\sum_{k=1}^{j-1} \phi(m_k, t_j - t_k)}$$



Pairwise influence, i.e. probability that v_j indirectly influences v_i:

$$r_{ij} = \sum_{k}^{j-1} r_{ik} p_{kj}$$

Influence of a $\varphi(v_i)$ tweet is then the sum of its pairwise influence:

$$\varphi(v_i) = \sum_{k=1}^{n} r_{ik}$$

Hawkes Processes for retweet cascades - Influence

Example - cascade of 224 retweets

- Magnitude is the number of followers a user has.
- Takes into consideration of when a user joins the Retweet cascade, and not only the number of followers.

time	magnitude	user_id	influence
0.000	1475.0	1169702293945630720	195.000000
621.928	142881.0	1090715513586679813	161.277176
565.046	16527.0	809115114	143.133077
165.381	2285.0	63003476	118.805366
304.526	591.0	1074137604851949569	27.393430
737.972	27081.0	770310096228417538	24.184821
1550.915	51256.0	824796324524498944	22.688434
544.080	546.0	18109811	20.321321

Hawkes Processes for retweet cascades - Influence

Example - cascade of 224 retweets

- Magnitude is the number of followers a user has.
- Takes into consideration of when a user joins the Retweet cascade, and not only the number of followers.

This method of getting user influence was implemented in and made easy-to-use in Scala using the same infrastructure that has been done for the rest of Twitter-data.

time	magnitude	user_id	influence
0.000	1475.0	1169702293945630720	195.000000
621.928	142881.0	1090715513586679813	161.277176
565.046	16527.0	809115114	143.133077
165.381	2285.0	63003476	118.805366
304.526	591.0	1074137604851949569	27.393430
737.972	27081.0	770310096228417538	24.184821
1550.915	51256.0	824796324524498944	22.688434
544.080	546.0	18109811	20.321321

Summary

I. Retweet cascades were modelled using marked Hawkes processes.
   A. This model can also be used for prediction active retweet cascades.
II. A way to estimate user influence in a retweet diffusion process has been implemented.
III. The first steps for being able to look at predictive causality between social media and mass media were taken.

See: https://github.com/lamastex/mep for public codes and https://github.com/lamastex/HawkesProcessesOnMedia for manuscript