Discover Twitter influential users by means of Markov chains models, a computational approach.

by *Riccardo Scalco, Sergio Cima*

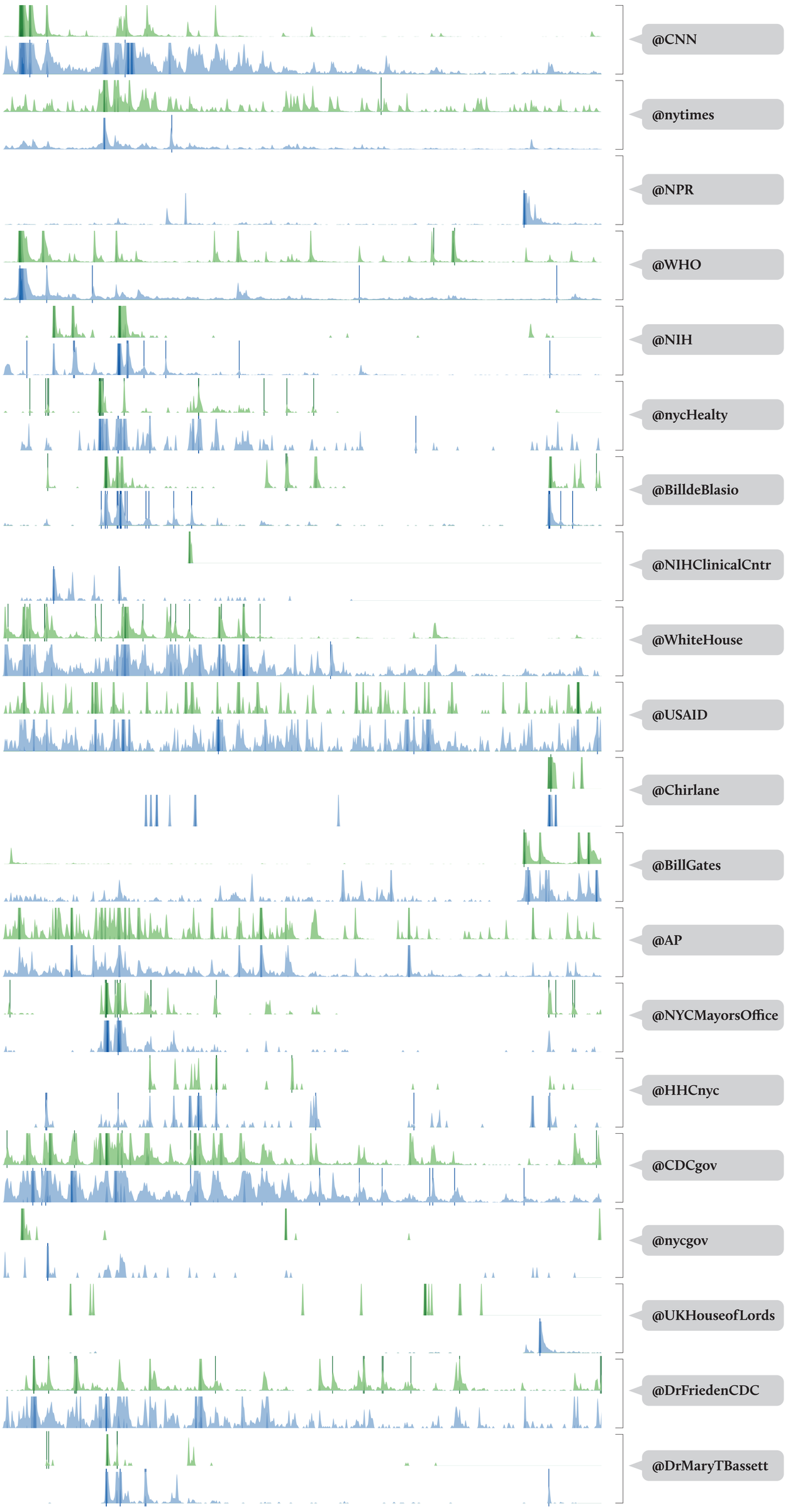

Fig. 1

The above figure shows the interaction events of the first **20 most influential users**. Two timeseries are associated to each user and both of them have a timeline that goes from **October 19** to **November 12**.
The two timeseries indicate the amount of times the user has been *retweeted* (green) or *mentioned* (blu) over time, with data integrated over a time step of 1 hour.
Vertical lines shown the presence of a *retweet/mention received by another influential user among the same list*, providing a view on network relationships over time.

Note how some of the massive events (the peaks) have at the very beginning an influential user (i.e. a vertical line).
Our assumption is that the **massive events can be triggered by such valuable events** (a further validation of such assumption is difficult because with the present state of Twitter API retweets of retweets do not show representations of the intermediary retweet, but only the original tweet).
The identification of the most valuable interacting users among all is a vital feature on the understanding of twitter cascades related to the activity of a given user.

Abstract:
We researched a method to define the most influential twitter users on a specific topic, we describe here the applied methodology and the results achieved. Briefly, tweets are collected via the ***twitter streaming API***, stored in sqlite databases and then processed in order to create a regular ***Markov chain***. The steady state distribution of the chain defines a metric on the set of twitter users, which can be used to retrieve an ordered list of users.

Data Retrieval:

Data has been retrieved from twitter, connecting to the twitter public streaming endpoint with the following request parameters: keywords tracked: ***"ebola"***, languages: ***"it, en, fr, de, es"***.
From the returned tweet objects, we selected and stored the following informations relative to each tweet: tweet id, language, text, date of creation, user id, username, mentions/hashtags/urls contained and, in case of retweets, the id and user id of the original tweet.
We collected ***6.786.843 tweets***, of which *3.739.603* are retweets and *4.580.781* contain at least one user mention.

Network definition:

The raw data retrieved from the streaming is a list of objects, each one containing information relative to a single tweet. From such data structure we derived a weighted directed graph by means of the following assumptions:

- ***graph nodes*** are twitter users and ***graph edges*** are links between twitter users
- if user a retweets user b then a ***directed link (a,b) is created***
- if user a mentions user b then a ***directed link (a,b) is created***
- each link (a,b) is ***weighted by the number of times it is found on the data***, i.e. by the number of times user a retweets or mentions user b.

The directed graph created with the previous described rules is disconnected and it has near ***10^5 nodes***. After some manipulations, the resulting strongly connected graph has ***653 nodes*** and ***17.381 edges***, with a ***ratio edges/nodes of 26.6***.

Markov Chain and the steady state

With an operation of normalization, the matrix of the link weights can be transformed into a ***right stochastic matrix T***. The so formed stochastic matrix T defines the transition matrix of a ***discrete-time finite Markov chain***.

The chain ***T*** is ergodic, which means that for any couple of nodes there exists a path connecting them with non zero probability. Ergodic chains do not present absorbing states, and the equilibrium distribution has non zero entries.

The steady state p, i.e. the equilibrium distribution, is unique and it is defined as the eigenvector of ***T*** with eigenvalue equal to one: ***p = pT***. The steady state distribution of the chain defines a metric on the set of twitter users, which can be used to retrieve an ordered list of users.

Observation:

The rules applied on the network definition have the target to highlight the user visibility on the discussion, we indeed preferred to focus on events that clearly determine a connection among users that is also visible to other users.

Being based on mentions and retweets, the described metodology is effective also on discovering influential users on the short period, therefore it offers an analytical tool suitable for discussions on specific events along with a time variable list of influential users.

The influence of a user is weighted by means of the influence of the users linked to herself, therefore mentions or retweet from influential users are weighted more.

Despite the similarities with the calculation of the eigenvector centrality, the resulting list of influential users is different. Eigenvector centrality takes into account only the graph topology and it is more appropriate on unweighted networks.

Eventually, the ordered list of influential users does not correlate with the absolute number of retweets or mentions achieved by the users, preventing not relevant but highly popular users to be qualified as influential.

A data research and analysis by:

TUX TAX

ZADIG
SALUTE, SCIENZA E AMBIENTE