Predicting persuadable individuals using text mining and graph machine learning on Twitter

Federico Albanese
(falbanese@dc.uba.ar)
@f__albanese

Esteban Feuerstein[1]

Leandro Lombardi[2]

Pablo Balenzuela[3]

[1] Instituto de Ciencias de la Computación (ICC), UBA-CONICET
[2] Instituto de Cálculo (IC), UBA-CONICET
[3] Instituto de Física de Buenos Aires (IFIBA), CONICET, Argentina.

"En el panóptico digital no existe ese Big Brother que nos extrae informaciones contra nuestra voluntad. Por el contrario, nos revelamos, incluso nos ponemos al desnudo por iniciativa propia."

Byung-Chul Han

Han, B. C. (2014). *Psicopolítica: neoliberalismo y nuevas técnicas de poder*. Herder Editorial.

"En el p *que nos*
ex *. Por el*
contra *desnudo*
propia."

Chul Han

Han, B. C. (2014). *Psicopolítica: neoliberalismo y nuevas técnicas de poder*. Herder Editorial.

Trabajos previos

Trabajos previos

- Conver et al. (2011): Usan detección de comunidades y muestran la estructura polarizada durante las elecciones del 2010 en Estados Unidos.

Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media

Political Polarization on Twitter

M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, F. Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington, IN, USA

Abstract

In this study we investigate how social media shape the networked public sphere and facilitate communication between communities with different political orientations. We examine two networks of political communication on Twitter, comprised of more than 250,000 tweets from the six weeks leading up to the 2010 U.S. congressional midterm elections. Using a combination of network clustering algorithms and manually-annotated data we demonstrate that the network of political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users. Surprisingly this is not the case for the user-to-user mention network, which is dominated by a single politically heterogeneous cluster of users in which ideologically-opposed individuals interact at a much higher rate compared to the network of retweets. To explain the distinct topologies of the retweet and mention networks we conjecture that politically motivated individuals provoke interaction by injecting partisan content into information streams whose primary audience consists of ideologically-opposed users. We conclude with statistical evidence in support of this hypothesis.

1 Introduction

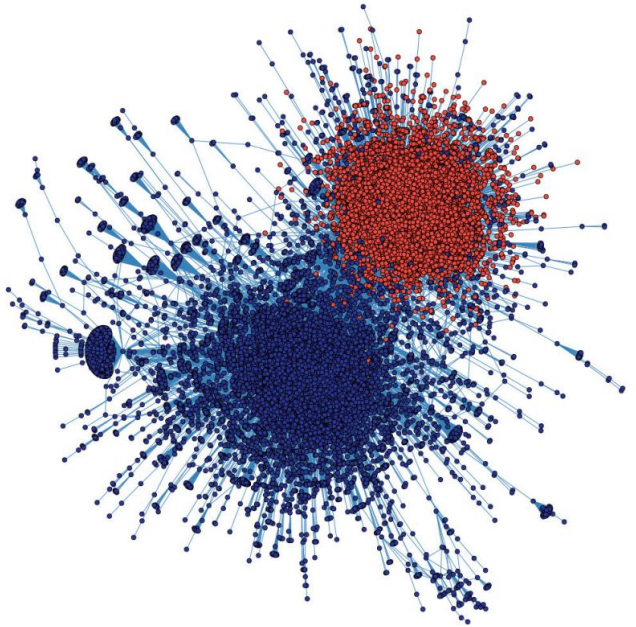
Social media play an important role in shaping political discourse in the U.S. and around the world (Bennett 2003; Benkler 2006; Sunstein 2007; Farrell and Drezner 2008; Aday et al. 2010; Tumasjan et al. 2010; O'Connor et al. 2010). According to the Pew Internet and American Life Project, six in ten U.S. internet users, nearly 44% of American adults, went online to get news or information about politics in 2008. Additionally, Americans are taking an active role in online political discourse, with 20% of internet users contributing comments or questions about the political process to social networking sites, blogs or other online forums (Pew Internet and American Life Project 2008). Gallo, and Kane (2007). Consumers of online political information tend to behave similarly, choosing to read blogs that share their political beliefs, with 26% more users doing so in 2008 than 2004 (Pew Internet and American Life Project 2008).

In its own right, the formation of online communities is not necessarily a serious problem. The concern is that when politically active individuals can avoid people and information they would not have chosen in advance, their opinions are likely to become increasingly extreme as a result of being exposed to more homogeneous viewpoints and fewer credible opposing opinions. The implications for the political process in this case are clear. A deliberative democracy relies on a broadly informed public and a healthy ecosystem of competing ideas. If individuals are exposed exclusively to people or facts that reinforce their pre-existing beliefs, democracy suffers (Sunstein 2002; 2007).

In this study we examine networks of political communication on the Twitter microblogging service during the six weeks prior to the 2010 U.S. midterm elections. Sampling data from the Twitter 'gardenhose' API, we identified 250,000 politically relevant messages (*tweets*) produced by more than 45,000 users. From these tweets we isolated two networks of political communication — the *retweet* network, in which users are connected if one has rebroadcast content produced by another, and the *mention* network, where users are connected if one has mentioned another in a post, including the case of tweet replies.

We demonstrate that the retweet network exhibits a highly modular structure, segregating users into two homogenous communities corresponding to the political left and right. In contrast, we find that the mention network does not exhibit this kind of political segregation, resulting in users being exposed to individuals and information they would not have been likely to choose in advance.

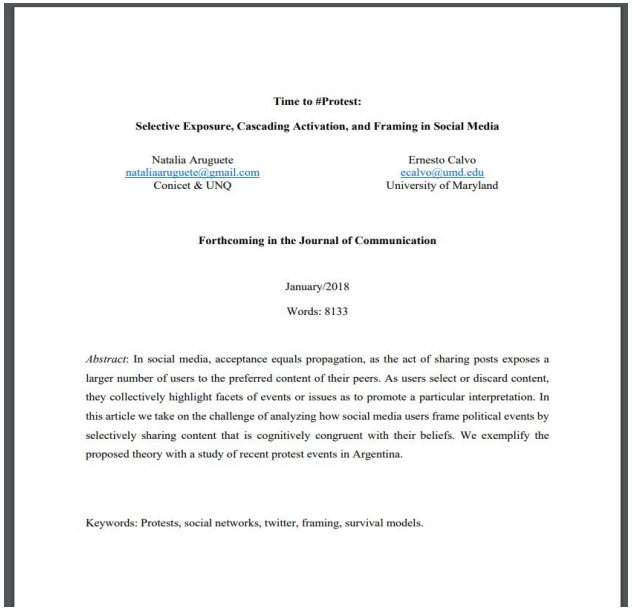
Finally, we provide evidence that these network structures



Red de Retweets (RT):
En **azul** Demócratas, en **rojo** Republicanos.

Trabajos previos

- Aruguete et al. (2018): muestran cómo los usuarios “like-minded” se agrupan e interactúan entre sí en la red de RT formando comunidades cerradas.



Time to #Protest:

Selective Exposure, Cascading Activation, and Framing in Social Media

Natalia Aruguete
nataliaaruguete@gmail.com
Conicet & UNQ

Ernesto Calvo
ecalvo@umd.edu
University of Maryland

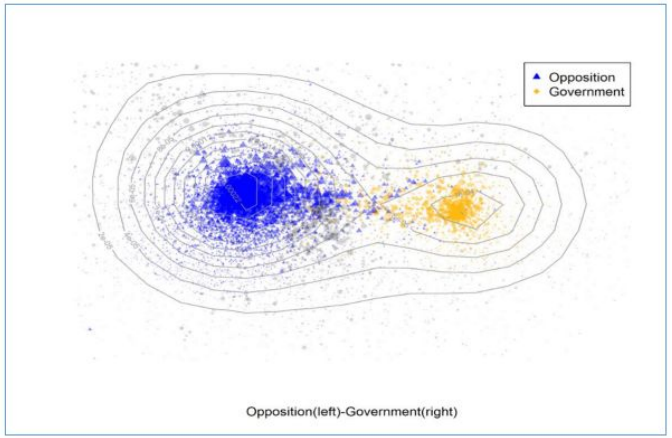
Forthcoming in the Journal of Communication

January/2018

Words: 8133

Abstract: In social media, acceptance equals propagation, as the act of sharing posts exposes a larger number of users to the preferred content of their peers. As users select or discard content, they collectively highlight facets of events or issues as to promote a particular interpretation. In this article we take on the challenge of analyzing how social media users frame political events by selectively sharing content that is cognitively congruent with their beliefs. We exemplify the proposed theory with a study of recent protest events in Argentina.

Keywords: Protests, social networks, twitter, framing, survival models.



Red de RT:
Azul: oposición; **Amarillo:** oficialismo; **Gris:** otros.

Trabajos previos

- Pinto et al. (2019): Detectan y analizan los tópicos de discusión para entender el rol de los medios de comunicación en el proceso de formación de opinión pública.

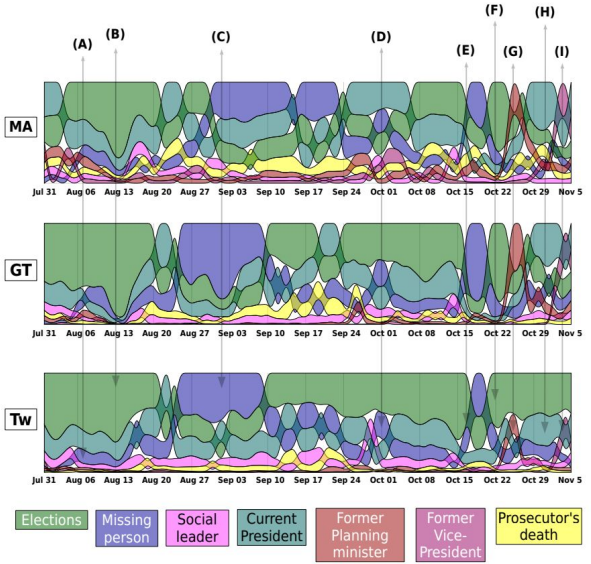
Quantifying time-dependent Media Agenda and Public Opinion by topic modeling

Sebastián Pinto *[1,2], Federico Albanese[3], Claudio O. Dorso[1,2], and Pablo Balenzuela[1,2]

[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Av.Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
[2]Instituto de Física de Buenos Aires (IFIBA), CONICET, Av.Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.
[3]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET, Av.Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428, Buenos Aires, Argentina.

Abstract

The mass media plays a fundamental role in the formation of public opinion, either by defining the topics of discussion or by making an emphasis on certain issues. Directly or indirectly, people get informed by consuming news from the media. Naturally, two questions appear: What are the dynamics of the agenda and how the people become interested in their different topics? These questions cannot be answered without proper quantitative measures of agenda dynamics and public attention. In this work we study the agenda of newspapers in comparison with public interests by performing topic detection over the news. We define Media Agenda as the distribution of topic's coverage by the newspapers and Public Agenda as the distribution of public interest in the same topic space. We measure agenda diversity as a function of time using the Shannon entropy and differences between agendas using the Jensen-Shannon distance. We found that the Public Agenda is less diverse than the Media Agenda, especially when there is a very attractive topic and the audience naturally focuses only on this one. Using the same methodology we detect coverage bias in newspapers. Finally, it was possible to identify a complex



Bump Graph de los temas:
MA: Medios, GT: Google Trends, Tw: twitter

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework
  - Objetivo

    Detectar ***cambios de agrupamiento político*** de las personas

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework
  - Objetivo

    Detectar ***cambios de agrupamiento político*** de las personas
  - Herramientas
    - Natural Language Processing techniques.
    - Graph Machine Learning.
    - Gradient Boosting Models.
  - Datos
    - Twitter

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework
  - Objetivo
    - Detectar cambios de agrupamiento político de las personas
  - Herramientas
    - Natural Language Processing techniques.
    - Graph Machine Learning.
    - Gradient Boosting Models.
  - Datos
    - Twitter

- Evaluamos el Framework

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework
  - Objetivo
    - Detectar cambios de agrupamiento político de las personas
  - Herramientas
    - Natural Language Processing techniques.
    - Graph Machine Learning.
    - Gradient Boosting Models.
  - Datos
    - Twitter

- Evaluamos el Framework

- Caracterizamos a los usuarios que cambian el agrupamiento político.
  - Análisis de feature importance.

¿Qué hicimos en este trabajo?

- Armamos un Machine Learning Framework
  - Objetivo
    - Detectar cambios de agrupamiento político de las personas
  - Herramientas
    - Natural Language Processing techniques.
    - Graph Machine Learning.
    - Gradient Boosting Models.
  - Datos
    - Twitter

- Evaluamos el Framework

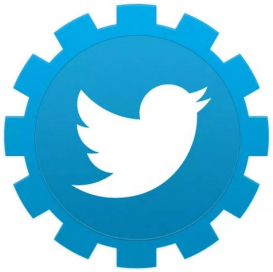
- Caracterizamos a los usuarios que cambian el agrupamiento político.
  - Análisis de feature importance.

- Identificación de los temas de discusión más importantes y persuasivos.

Data Collection

Data Collection

- ¿Dónde?
  Argentina.
- ¿En qué idioma?
  Español.
- ¿Cuándo?
  Durante la semana de las elecciones.
- ¿Queries?
  Nombres de candidatos, partidos y políticos.

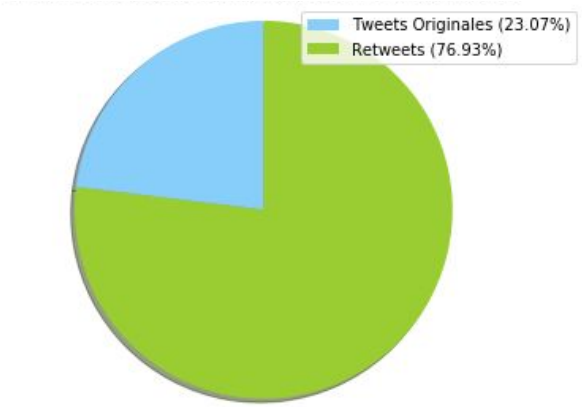


Streaming API

Datos

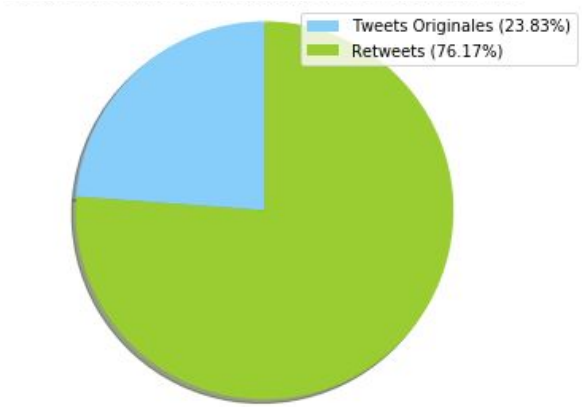
Elecciones PASO 2017:

- 2.117.708 tweets
- 1.629.200 RT
- 86.361 usuarios



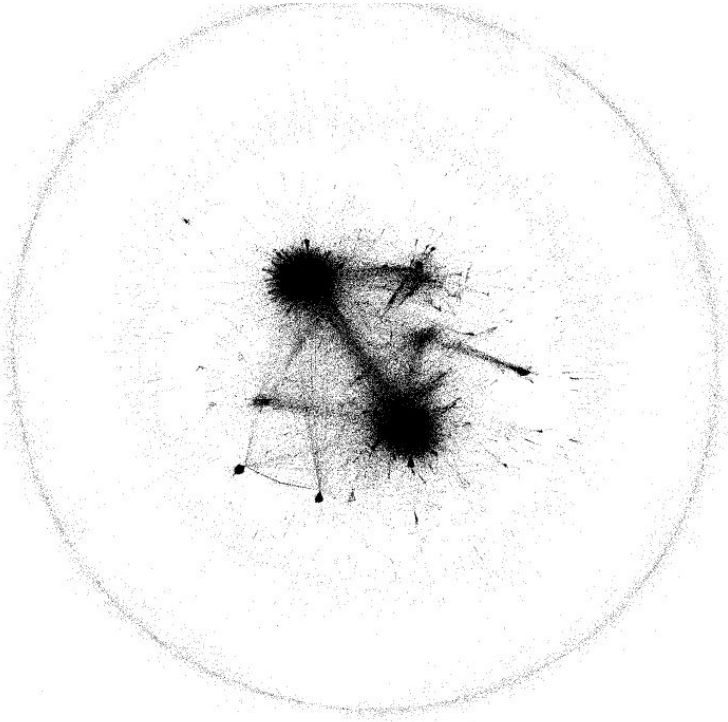
Elecciones Generales 2017:

- 1.751.813 tweets
- 1.334.294 RT
- 298.866 usuarios



Análisis de los datos

Red de RT

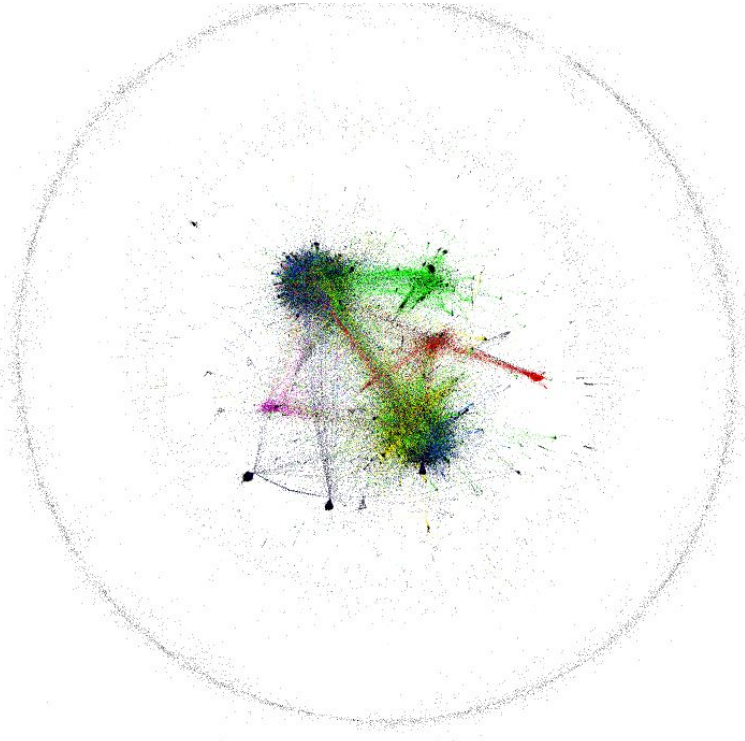


Visualización hecha con *Force Atlas 2* con los datos de las elecciones PASO 2017.

Red de RT

¿De quién hablan?

- ***Unidad Ciudadana***
- ***Cambiemos***
- ***Frente Justicialista***
- ***1Pais***
- ***Macri***

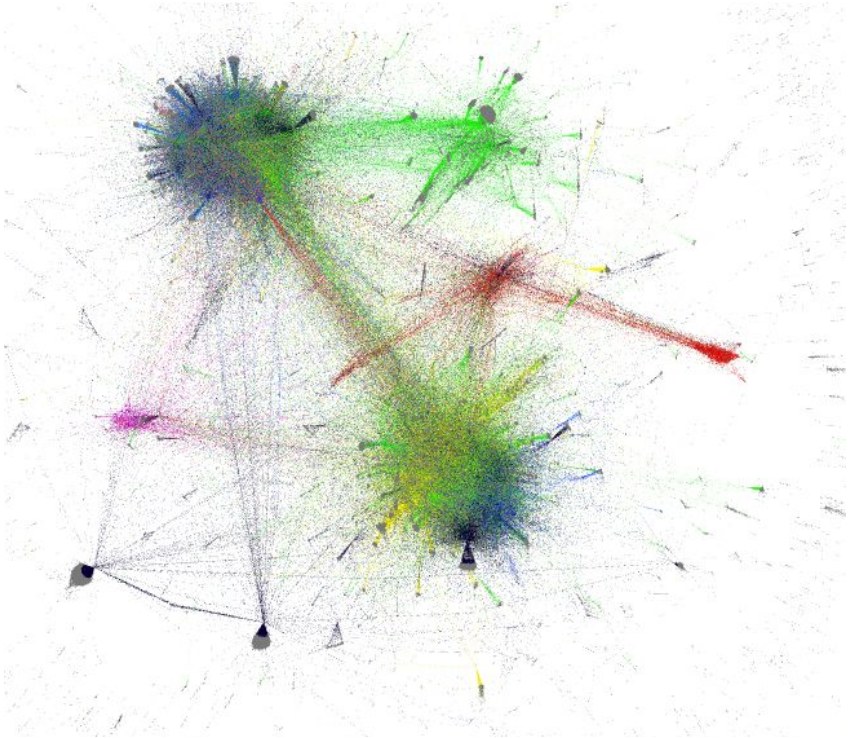


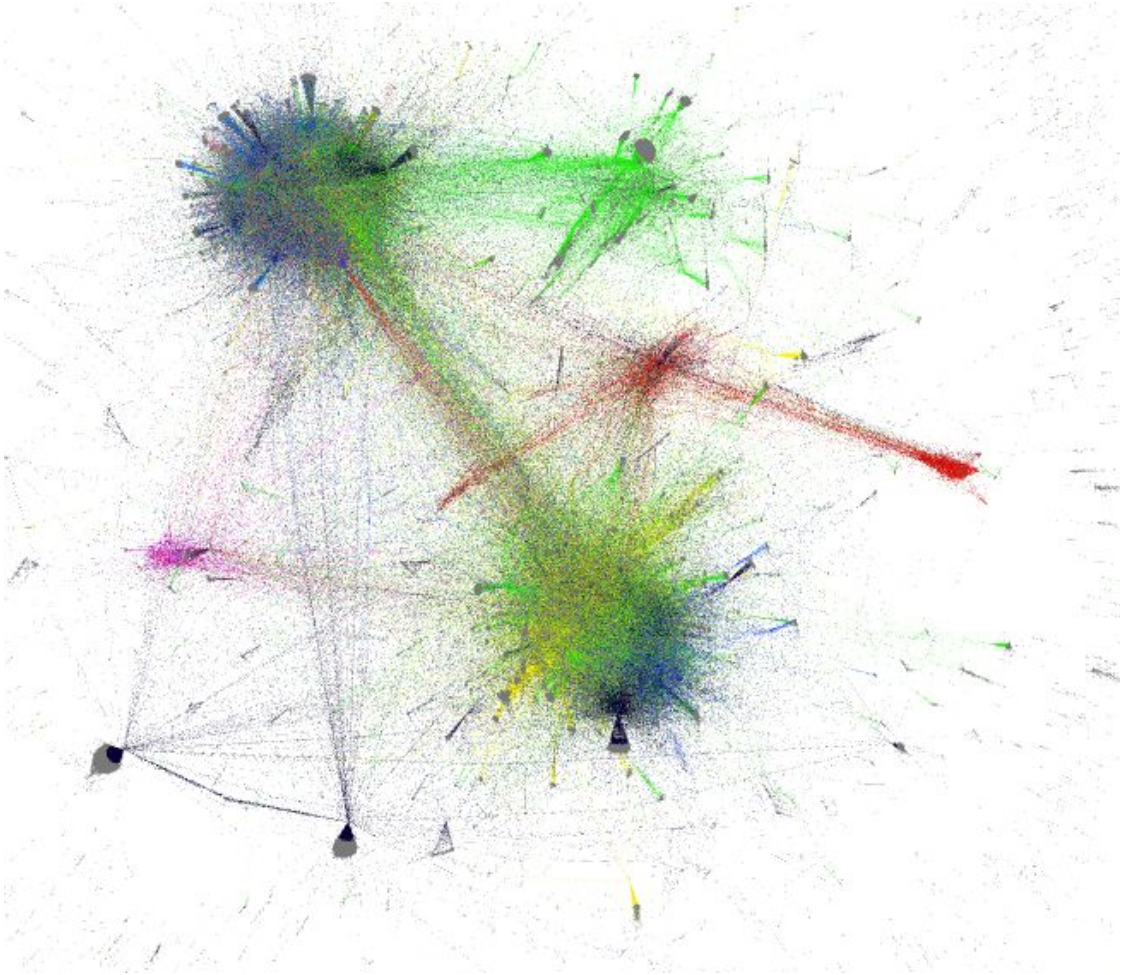
Visualización hecha con *Force Atlas 2* con los datos de las elecciones PASO 2017.

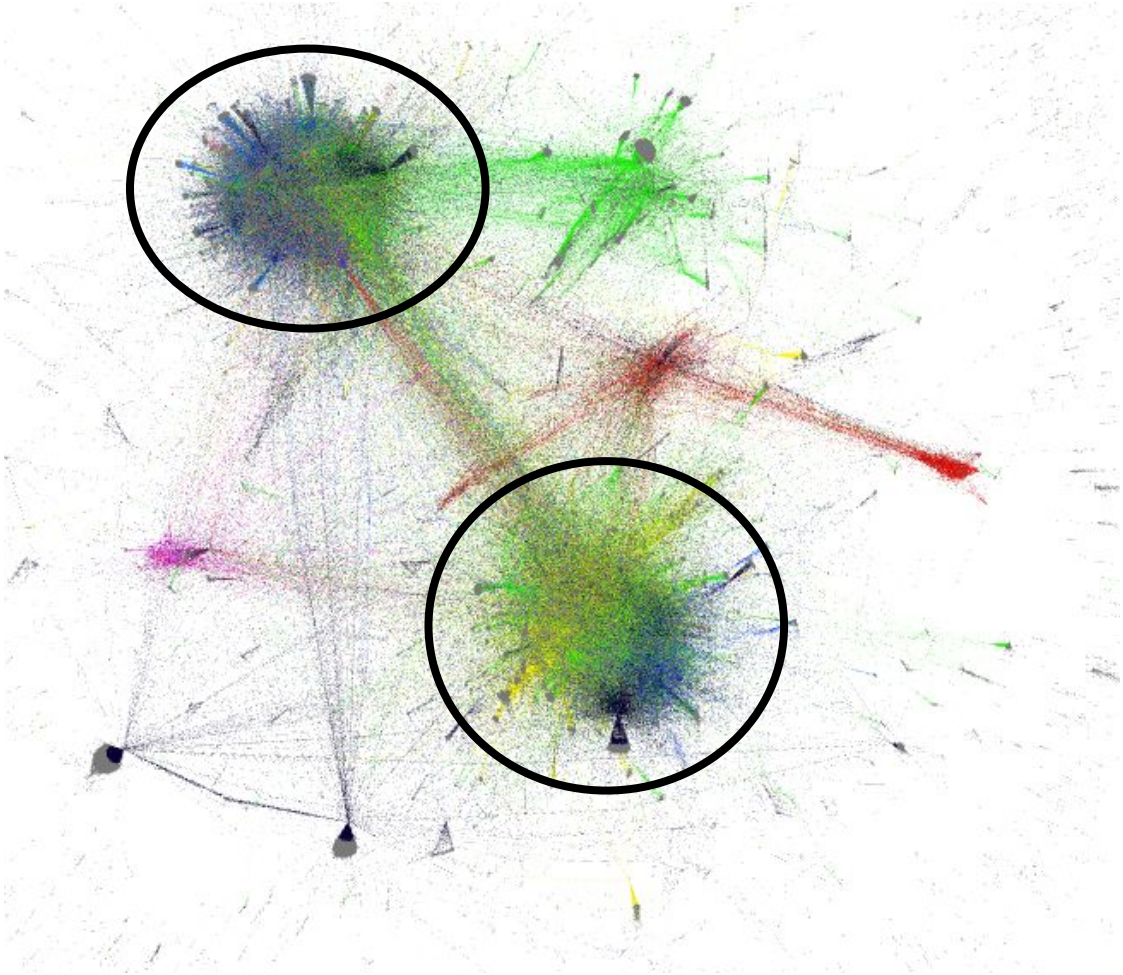
Red de RT

¿De quién hablan?

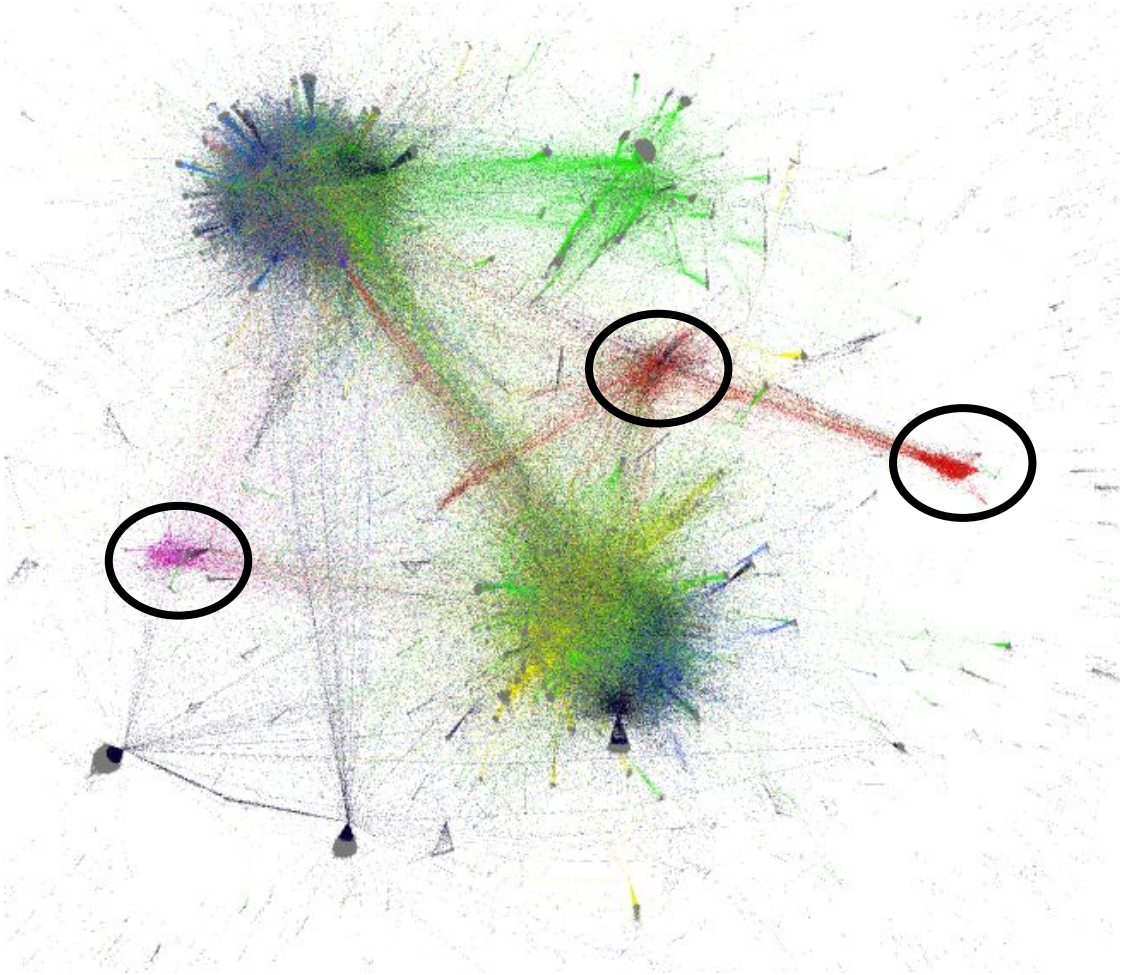
- ***Unidad Ciudadana***
- ***Cambiemos***
- ***Frente Justicialista***
- ***1Pais***
- ***Macri***



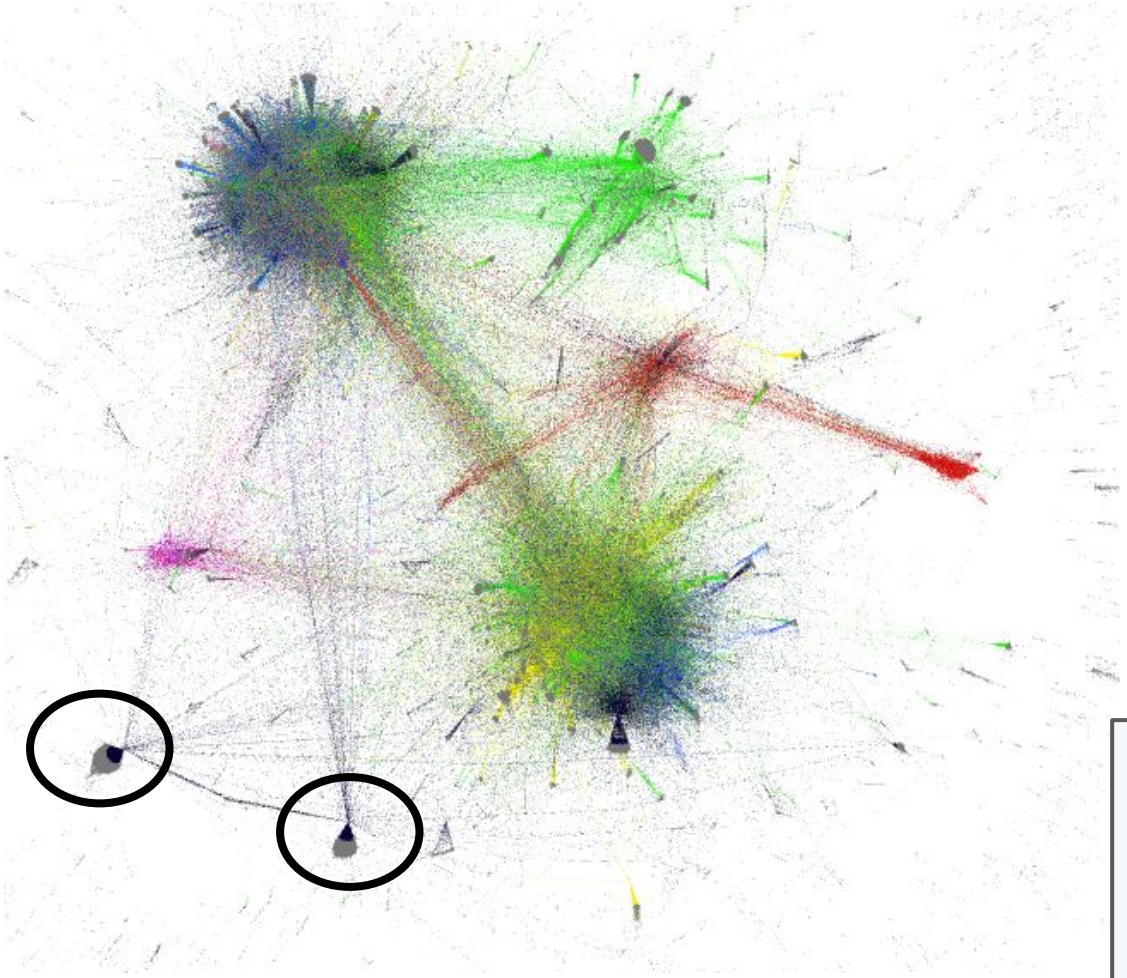




- Dos comunidades principales:
  “la grieta”



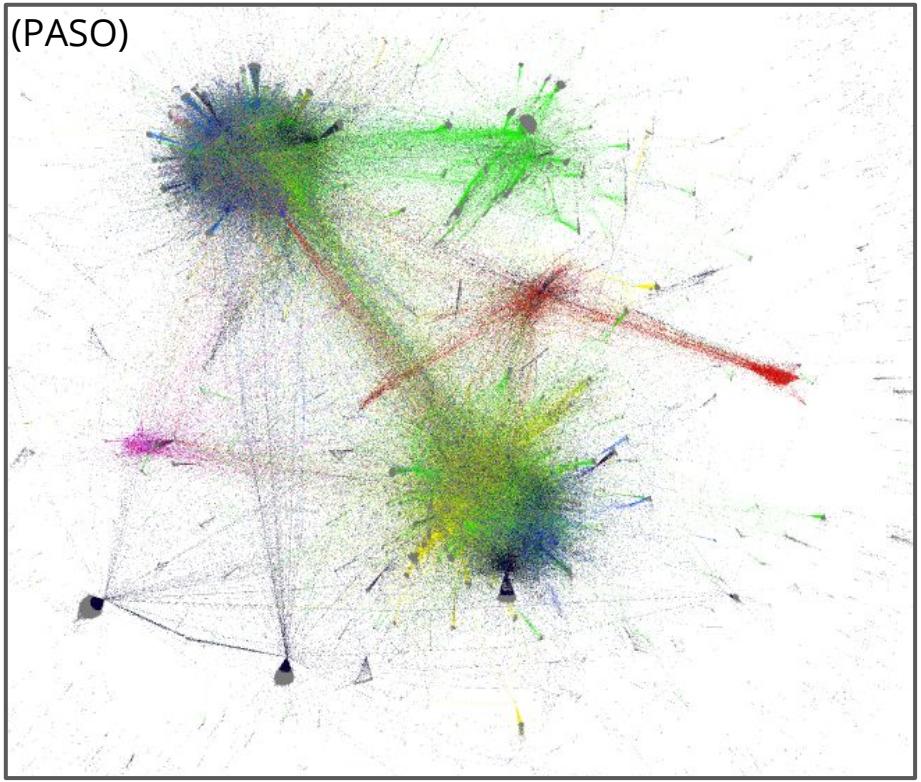
- Dos comunidades principales: ***“la grieta”.***
- Comunidades pequeñas de los otros partidos.



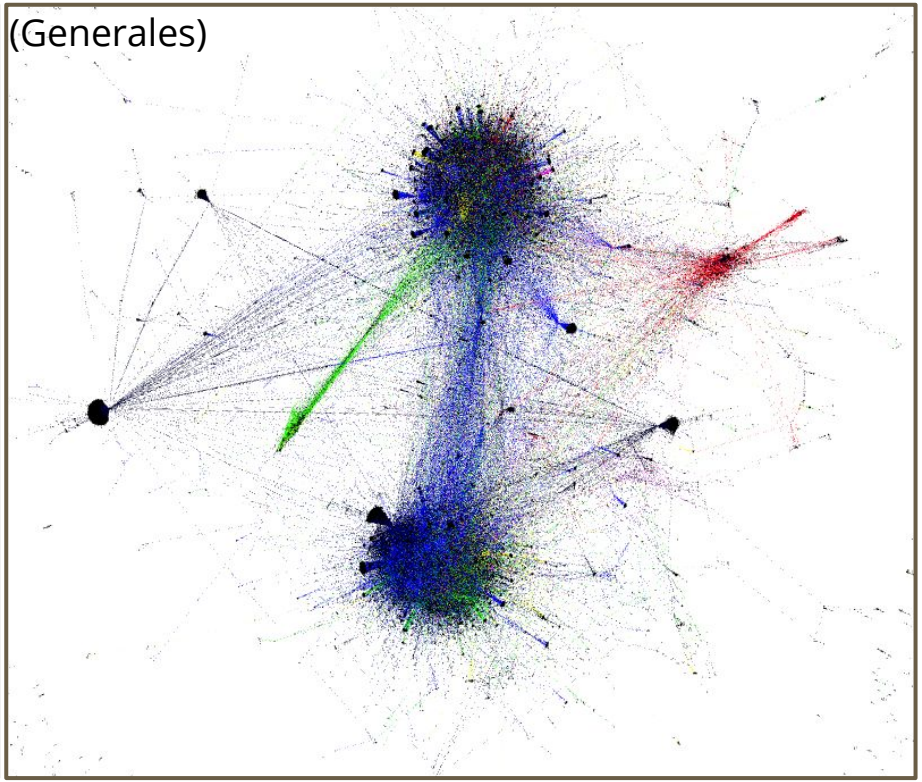
- Dos comunidades principales: ***“la grieta”.***
- Comunidades pequeñas de los otros partidos.
- Famosos por un dia.

Redes de RT

(PASO)

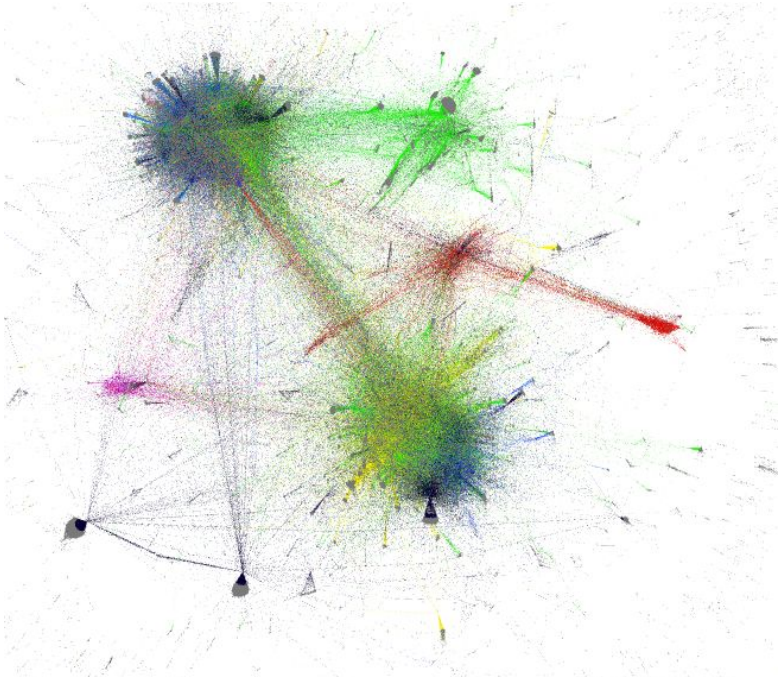


(Generales)

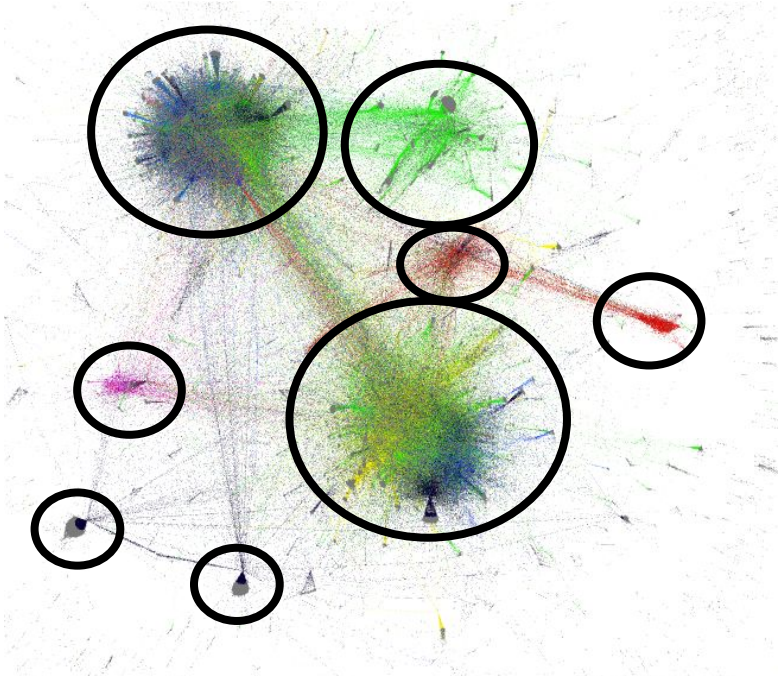


Unsupervised Machine Learning: Clustering

Unsupervised Machine Learning: Clustering



Unsupervised Machine Learning: Clustering



Método de Louvain

1. A todo nodo en la red se le asigna a una comunidad.

Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)

Método de Louvain

1. A todo nodo en la red se le asigna a una comunidad.

2. Para cada nodo, se le cambia iterativamente la comunidad por la de un vecino. Uno se queda con el cambio que más aumenta la modularidad.

Modularidad:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

Donde:

- A es la matriz de adyacencia.
- k_i es la suma de los pesos de las aristas del nodo i
- 2m es la suma de los pesos de todas las aristas de la red.
- c_i es la comunidad del nodo i.

Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)

Método de Louvain

1. A todo nodo en la red se le asigna a una comunidad.

2. Para cada nodo, se le cambia iterativamente la comunidad por la de un vecino. Uno se queda con el cambio que más aumenta la modularidad.

3. Se repite el paso iterativo hasta llegar al máximo de modularidad.

Modularidad:

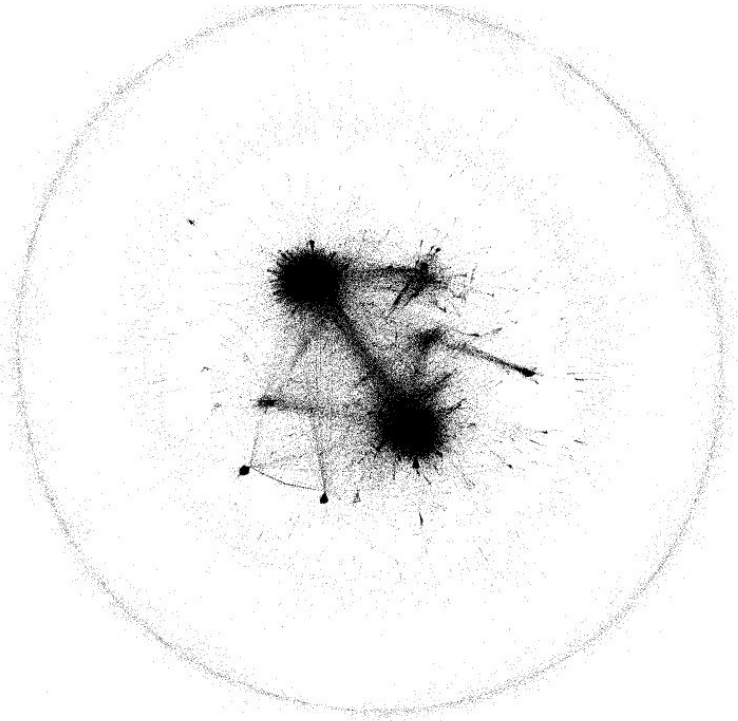
$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

Donde:

- A es la matriz de adyacencia.
- k_i es la suma de los pesos de las aristas del nodo i
- 2m es la suma de los pesos de todas las aristas de la red.
- c_i es la comunidad del nodo i.

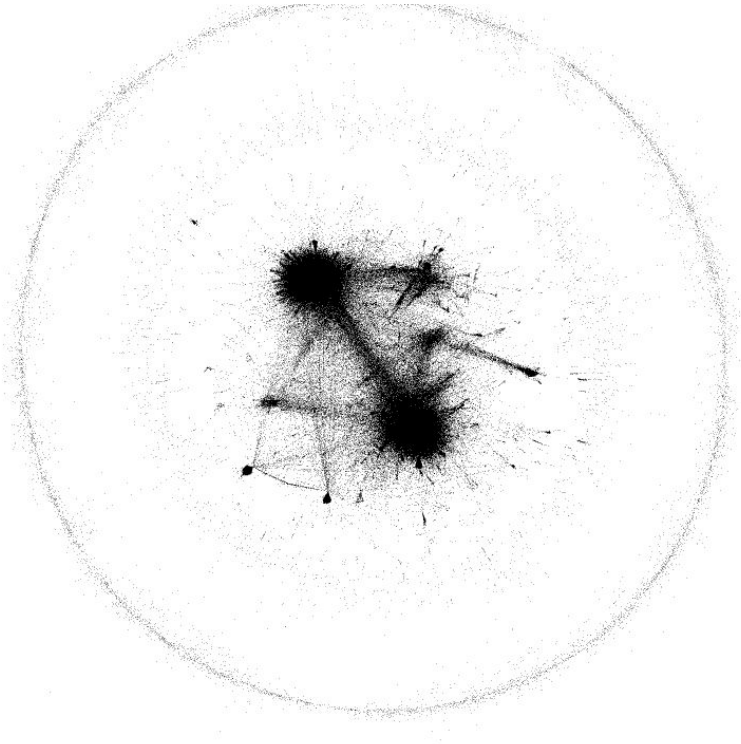
Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)

Graph Features



Graph Features

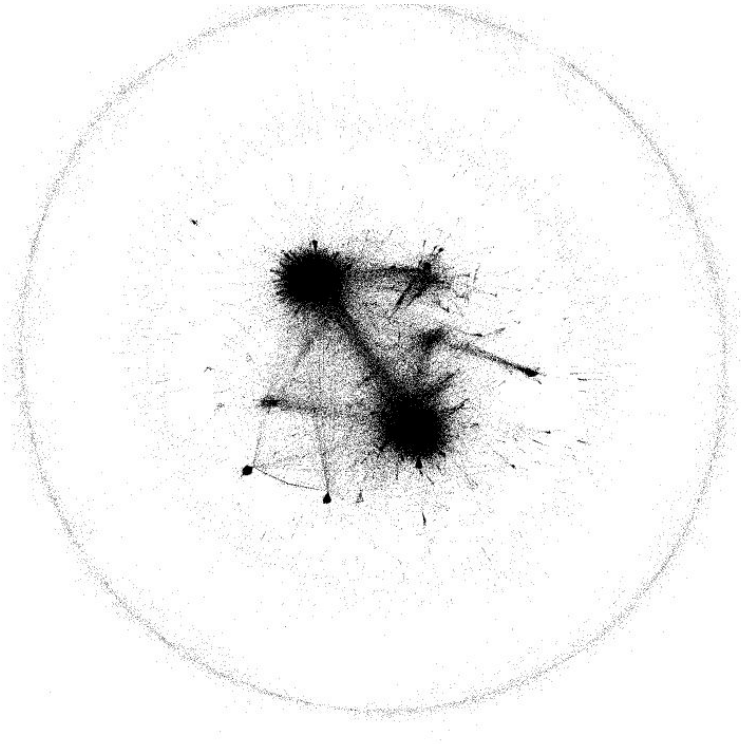
- Degree



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

Graph Features

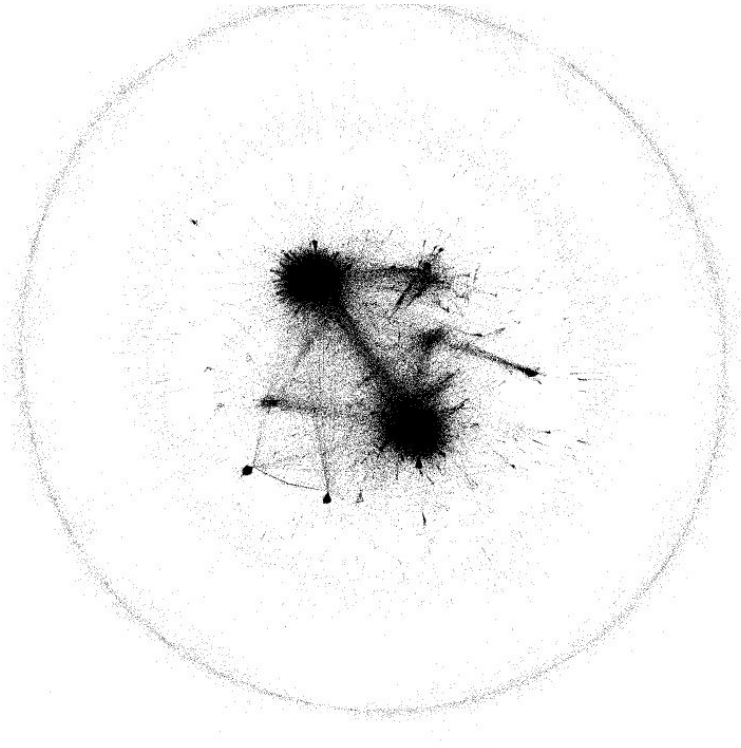
- Degree
- PageRank [1]



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

Graph Features

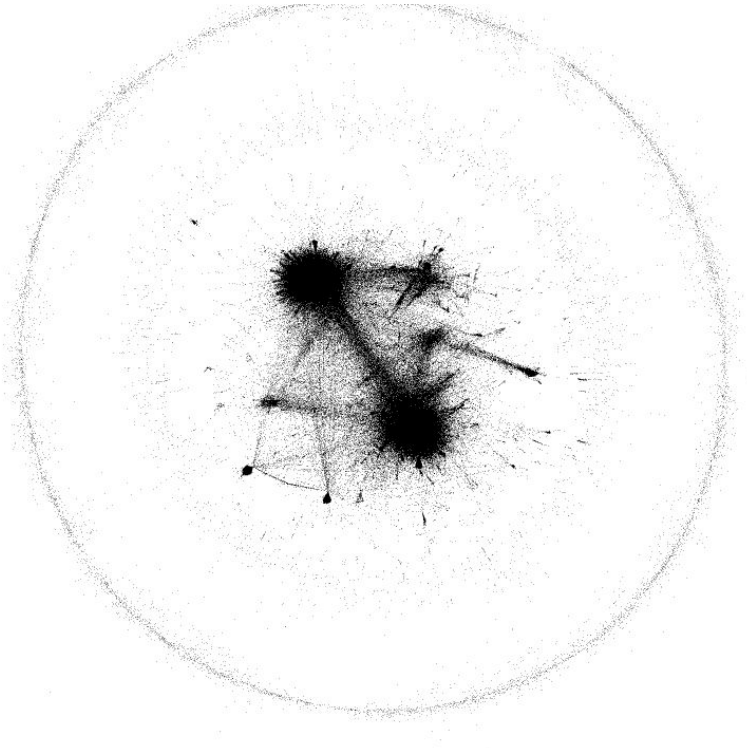
- Degree
- PageRank [1]
- Betweenness centrality [2]



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

Graph Features

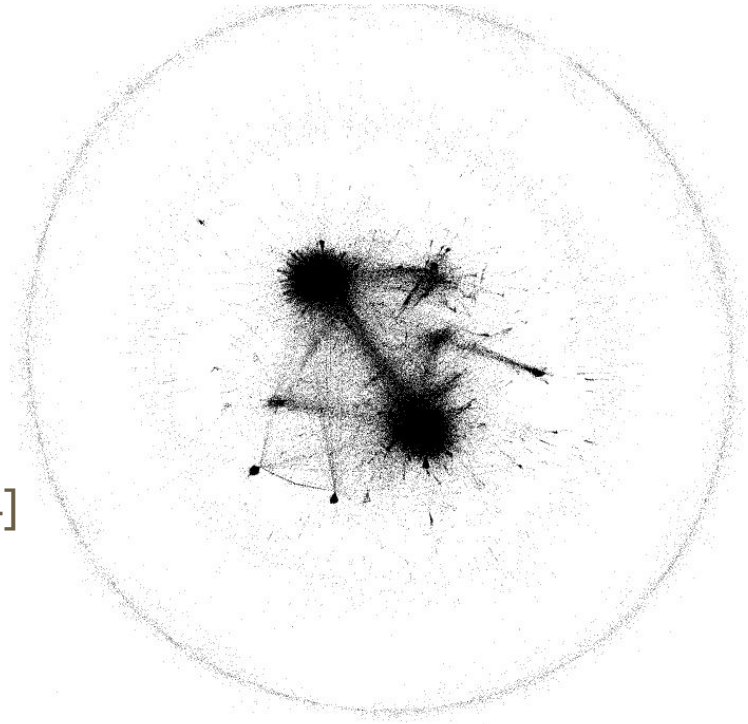
- Degree
- PageRank [1]
- Betweenness centrality [2]
- Clustering coefficient [3]



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

Graph Features

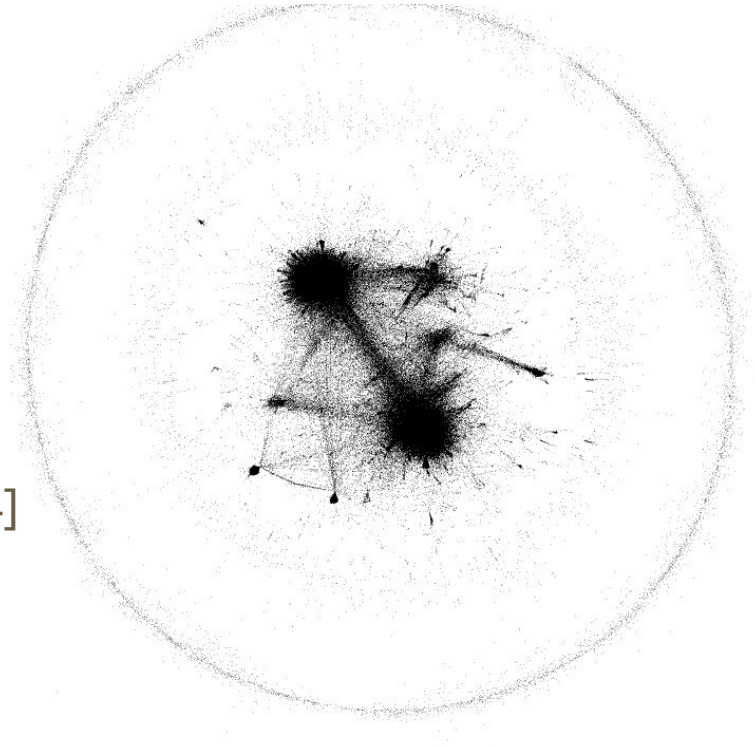
- Degree
- PageRank [1]
- Betweenness centrality [2]
- Clustering coefficient [3]
- Clustering affiliation (Louvain) [4]



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

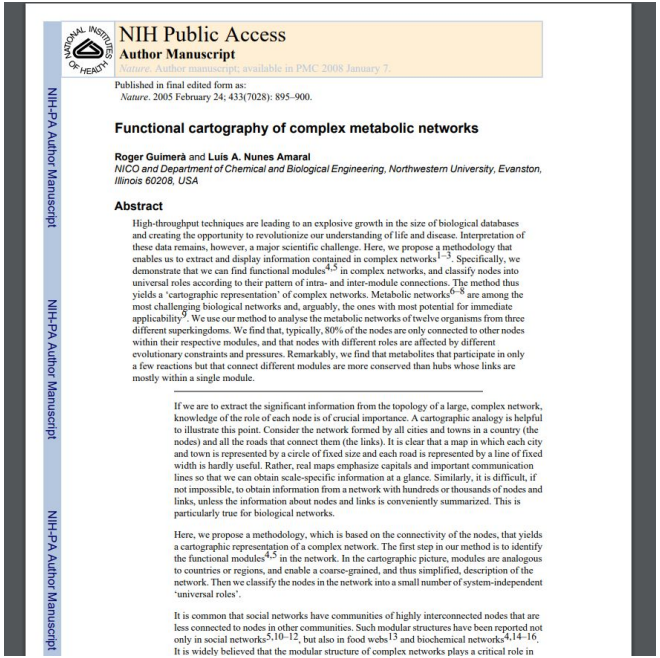
Graph Features

- Degree
- PageRank [1]
- Betweenness centrality [2]
- Clustering coefficient [3]
- Clustering affiliation (Louvain) [4]
- z-score del grado interno [5]
- Coeficiente de participación [5]



[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab
[2] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, 25(2), 163-177.
[3] Saramki, J., Kivel, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.
[4] Fast unfolding of communities in large networks, Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
[5] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895.

Graph Features: cartografia de la red



NIH Public Access
Author Manuscript
Nature. Author manuscript; available in PMC 2008 January 7.

Published in final edited form as:
Nature. 2005 February 24; 433(7028): 895–900.

Functional cartography of complex metabolic networks

Roger Guimerà and **Luis A. Nunes Amaral**
NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

Abstract

High-throughput techniques are leading to an explosive growth in the size of biological databases and creating the opportunity to revolutionize our understanding of life and disease. Interpretation of these data remains, however, a major scientific challenge. Here, we propose a methodology that enables us to extract and display information contained in complex networks[1–3]. Specifically, we demonstrate that we can find functional modules[4,5] in complex networks, and classify nodes into universal roles according to their pattern of intra- and inter-module connections. The method thus yields a 'cartographic representation' of complex networks. Metabolic networks[6–8] are among the most challenging biological networks and, arguably, the ones with most potential for immediate applicability[9]. We use our method to analyse the metabolic networks of twelve organisms from three different superkingdoms. We find that, typically, 80% of the nodes are only connected to other nodes within their respective modules, and that nodes with different roles are affected by different evolutionary constraints and pressures. Remarkably, we find that metabolites that participate in only a few reactions but that connect different modules are more conserved than hubs whose links are mostly within a single module.

If we are to extract the significant information from the topology of a large, complex network, knowledge of the role of each node is of crucial importance. A cartographic analogy is helpful to illustrate this point. Consider the network formed by all cities and towns in a country (the nodes) and all the roads that connect them (the links). It is clear that a map in which each city and town is represented by a circle of fixed size and each road is represented by a line of fixed width is hardly useful. Rather, real maps emphasize capitals and important communication lines so that we can obtain scale-specific information at a glance. Similarly, it is difficult, if not impossible, to obtain information from a network with hundreds or thousands of nodes and links, unless the information about nodes and links is conveniently summarized. This is particularly true for biological networks.

Here, we propose a methodology, which is based on the connectivity of the nodes, that yields a cartographic representation of a complex network. The first step in our method is to identify the functional modules[4,5] in the network. In the cartographic picture, modules are analogous to countries or regions, and enable a coarse-grained, and thus simplified, description of the network. Then we classify the nodes in the network into a small number of system-independent 'universal roles'.

It is common that social networks have communities of highly interconnected nodes that are less connected to nodes in other communities. Such modular structures have been reported not only in social networks[5,10–12], but also in food webs[13] and biochemical networks[4,14–16]. It is widely believed that the modular structure of complex networks plays a critical role in

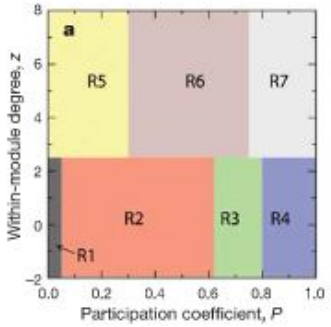
NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

- z-score del grado interno:

$$Z_i = \frac{(K_i - K_s)}{S_s}$$

- Coeficiente de participación:

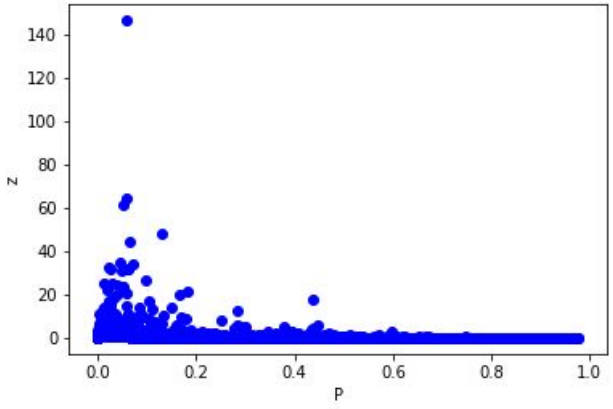
$$P_i = 1 - \sum_s \left(K_{i,s} - K_i\right)^2$$



[1] Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, *76*(3), 036106.
[2] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *nature*, *433*(7028), 895.

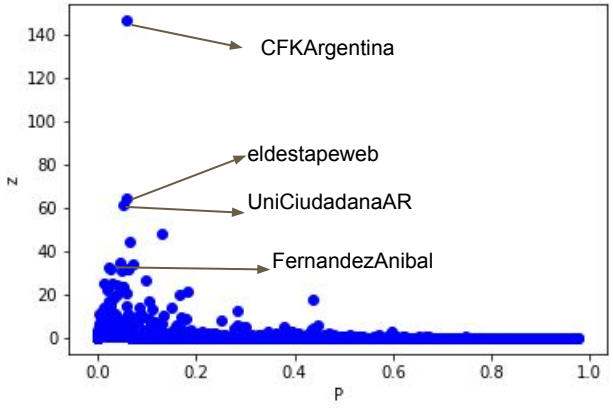
Graph Features: cartografia de la red

Comunidad:
Unidad Ciudadana



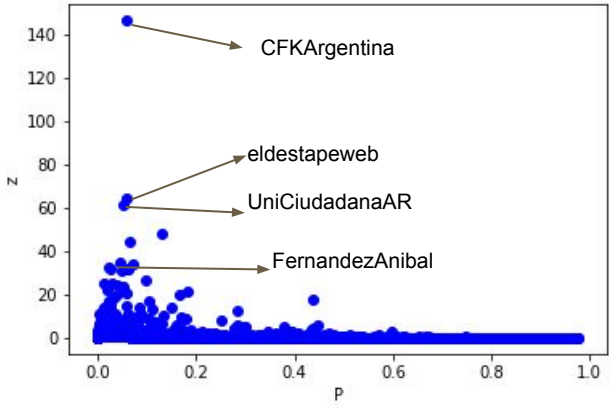
Graph Features: cartografia de la red

Comunidad:
Unidad Ciudadana

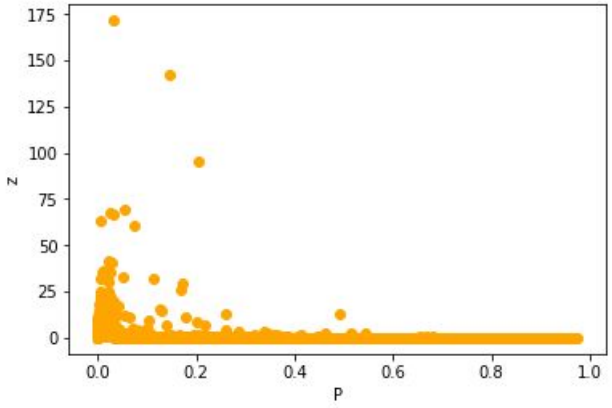


Graph Features: cartografia de la red

Comunidad:
Unidad Ciudadana

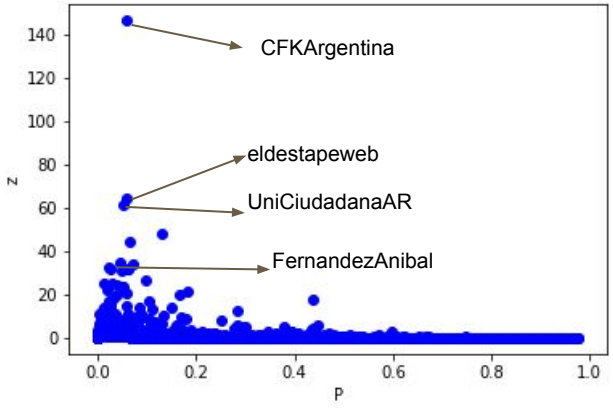


Comunidad:
Cambiemos

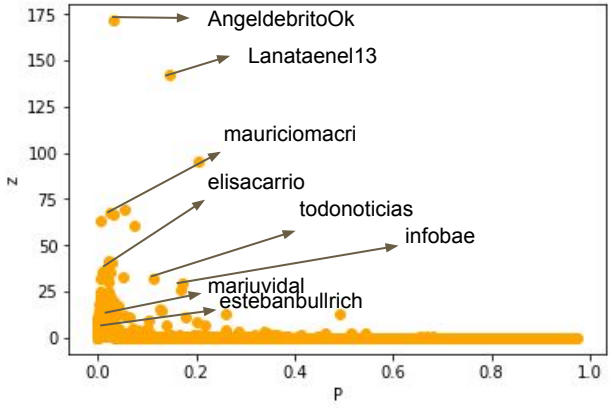


Graph Features: cartografia de la red

Comunidad:
Unidad Ciudadana

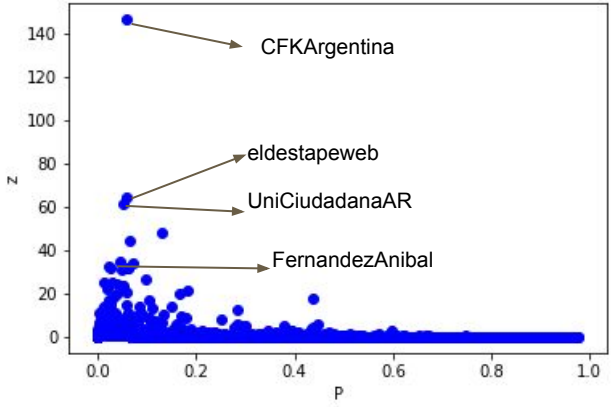


Comunidad:
Cambiemos

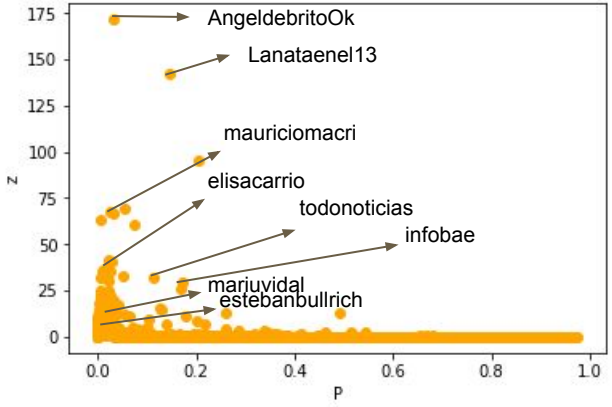


Graph Features: cartografia de la red

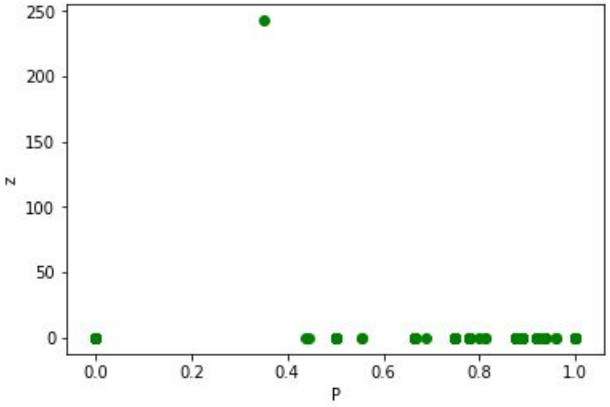
Comunidad:
Unidad Ciudadana



Comunidad:
Cambiemos

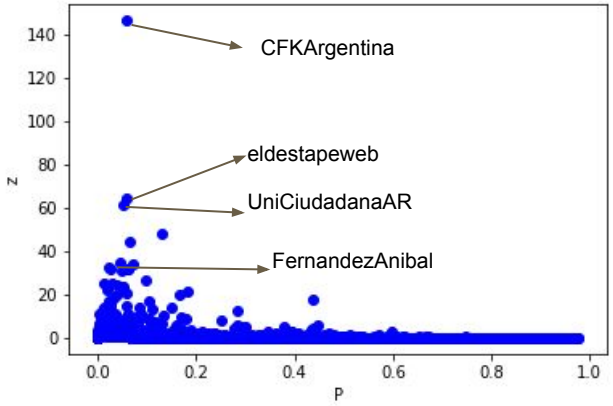


Comunidad:
Famoso por un dia

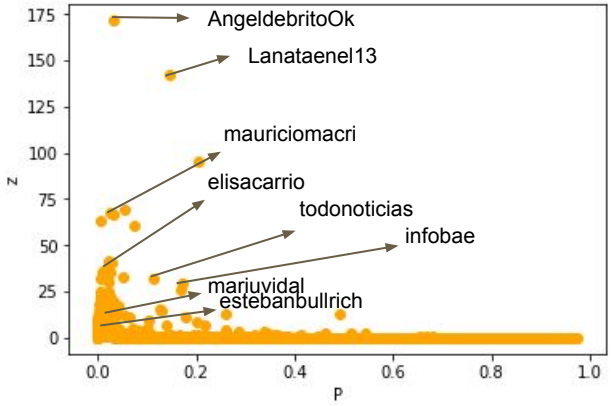


Graph Features: cartografia de la red

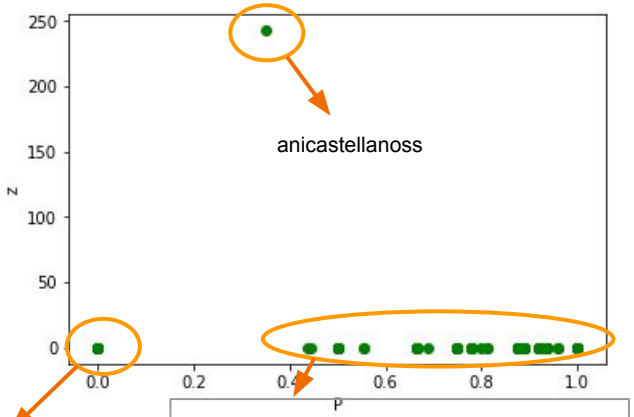
Comunidad:
Unidad Ciudadana



Comunidad:
Cambiemos



Comunidad:
Famoso por un dia



Usuarios con un único RT a anicastellanoss (z=0 y p=0)

A diferencia de los otros casos, este cluster no está definido ideológicamente y la mayoría de los usuarios están muy conectados a otros clusters (p>0.4) pero poco dentro del mismo cluster (z=0)

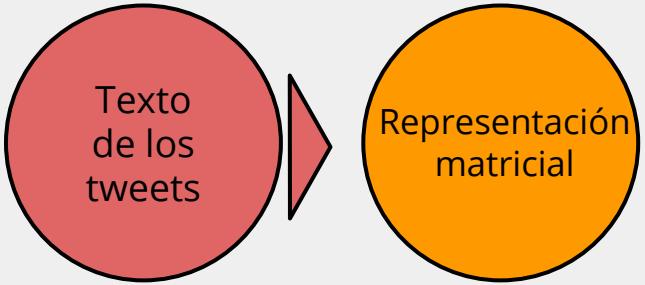
Text mining

Text mining

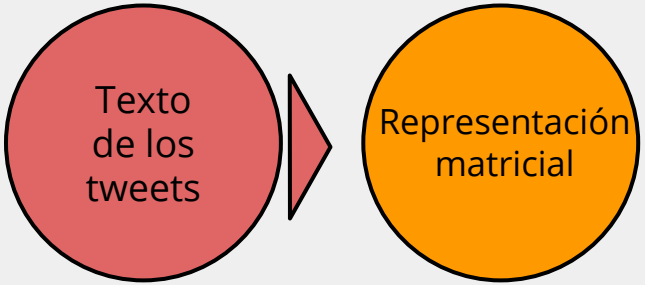
Text mining

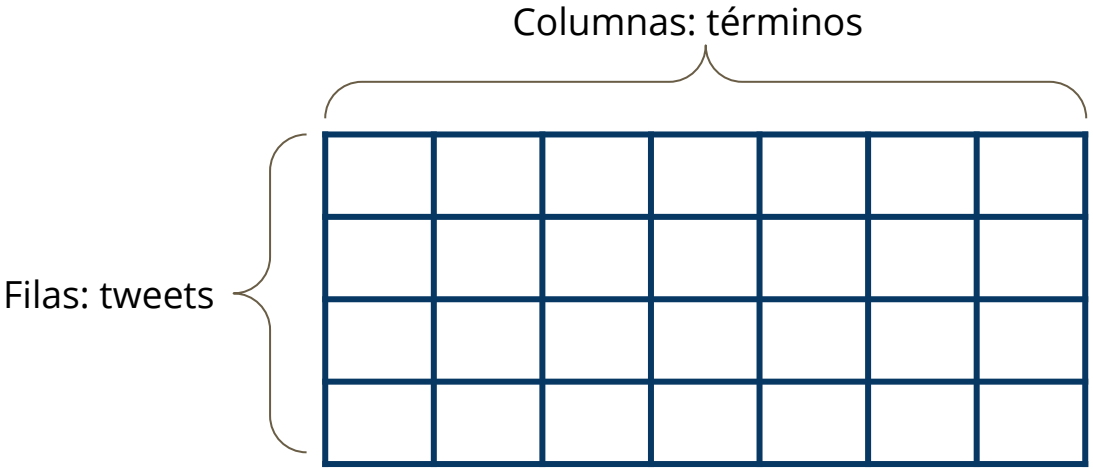
Texto de los tweets

Text mining

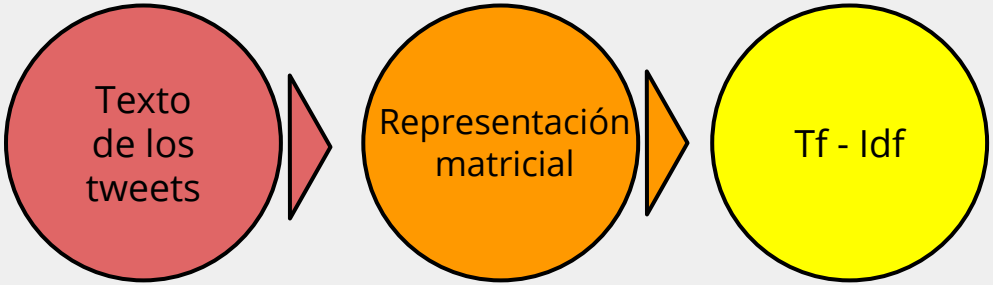


Text mining

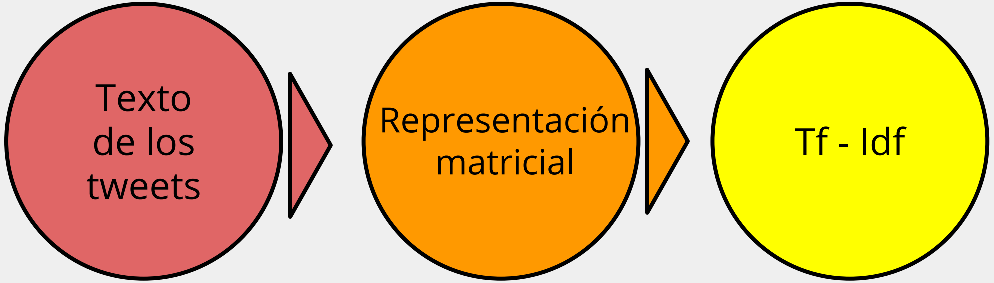

Texto de los tweets
Representación matricial

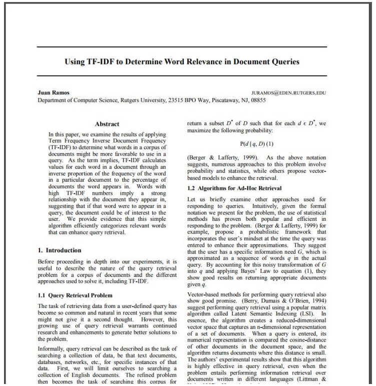

Columnas: términos
Filas: tweets

Text mining



Text mining





Using TF-IDF to Determine Word Relevance in Document Queries

Juan Ramos JURAMOS@EDEN.RUTGERS.EDU
Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855

Abstract

In this paper, we examine the results of applying Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a corpus of documents might be more favorable to use in a query. As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. We provide evidence that this simple algorithm efficiently categorizes relevant words that can enhance query retrieval.

1. Introduction

Before proceeding in depth into our experiments, it is useful to describe the nature of the query retrieval problem for a corpus of documents and the different approaches used to solve it, including TF-IDF.

1.1 Query Retrieval Problem

The task of retrieving data from a user-defined query has become so common and natural in recent years that some might not give it a second thought. However, this growing use of query retrieval warrants continued research and enhancements to generate better solutions to the problem.

Informally, query retrieval can be described as the task of searching a collection of data, be that text documents, databases, networks, etc., for specific instances of that data. First, we will limit ourselves to searching a collection of English documents. The refined problem then becomes the task of searching this corpus for return a subset D^* of D such that for each $d \in D^*$, we maximize the following probability:

$$P(d \mid q, D) \quad (1)$$

(Berger & Lafferty, 1999). As the above notation suggests, numerous approaches to this problem involve probability and statistics, while others propose vector-based models to enhance the retrieval.

1.2 Algorithms for Ad-Hoc Retrieval

Let us briefly examine other approaches used for responding to queries. Intuitively, given the formal notation we present for the problem, the use of statistical methods has proven both popular and efficient in responding to the problem. (Berger & Lafferty, 1999) for example, propose a probabilistic framework that incorporates the user's mindset at the time the query was entered to enhance their approximations. They suggest that the user has a specific information need G, which is approximated as a sequence of words q in the actual query. By accounting for this noisy transformation of G into q and applying Bayes' Law to equation (1), they show good results on returning appropriate documents given q.

Vector-based methods for performing query retrieval also show good promise. (Berry, Dumais & O'Brien, 1994) suggest performing query retrieval using a popular matrix algorithm called Latent Semantic Indexing (LSI). In essence, the algorithm creates a reduced-dimensional vector space that captures an n-dimensional representation of a set of documents. When a query is entered, its numerical representation is compared the cosine-distance of other documents in the document space, and the algorithm returns documents where this distance is small. The authors' experimental results show that this algorithm is highly effective in query retrieval, even when the problem entails performing information retrieval over documents written in different languages (Littman &

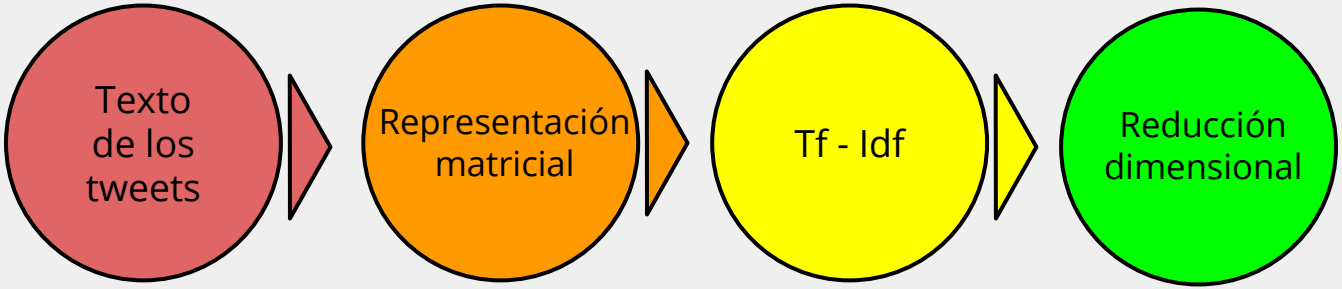
$$TF = N_t$$

$$IDF = 1 + log(\frac{1+N}{1+n_t})$$

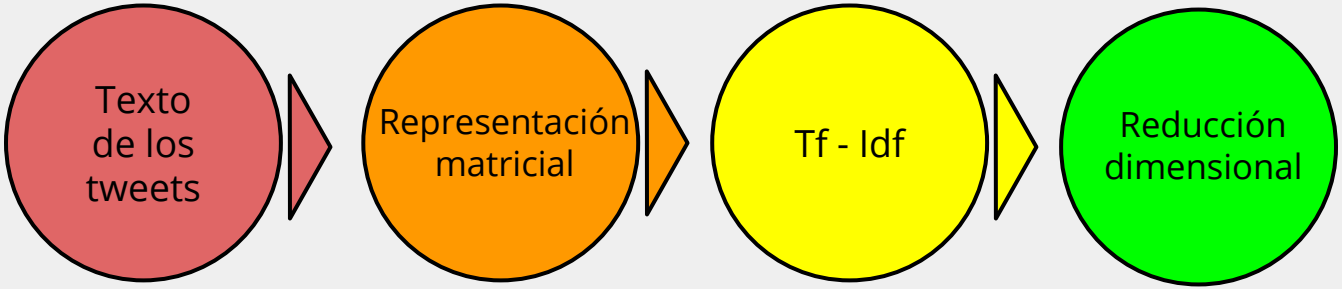
Donde N es el # total de documentos y N_t el # de documentos donde aparece el término t.

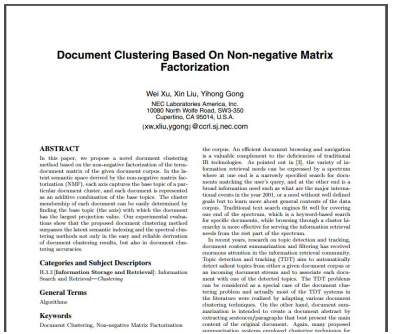
Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142).

Text mining


Texto de los tweets
Representación matricial
Tf - Idf
Reducción dimensional

Text mining





Document Clustering Based On Non-negative Matrix Factorization

Wei Xu, Xin Liu, Yihong Gong

NEC Laboratories America, Inc.
10080 North Wolfe Road, SW3-350
Cupertino, CA 95014, U.S.A.
{xw,xliu,ygong}@ccrl.sj.nec.com

ABSTRACT

In this paper, we propose a novel document clustering method based on the non-negative factorization of the term-document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. Our experimental evaluations show that the proposed document clustering method surpasses the latest semantic indexing and the spectral clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies.

Categories and Subject Descriptors

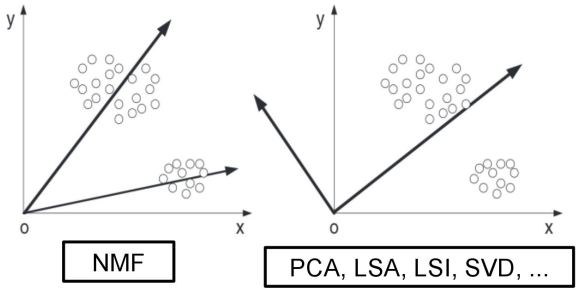
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms

Keywords

Document Clustering, Non-negative Matrix Factorization

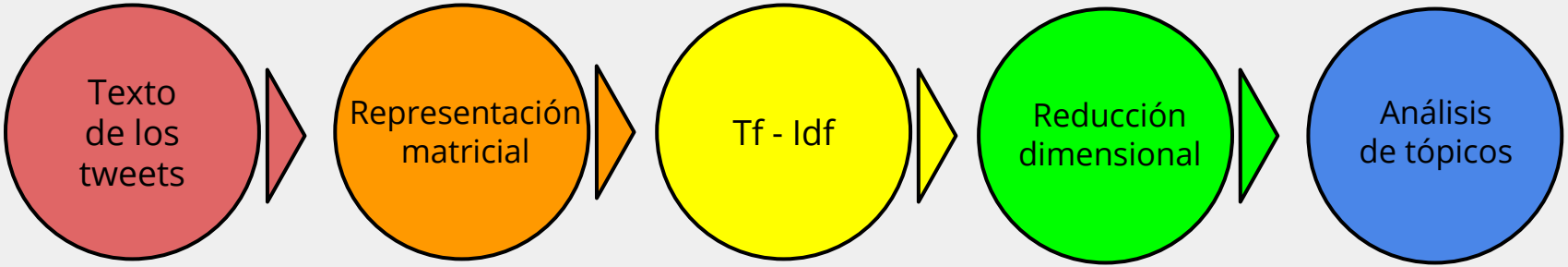


$$M \simeq H * W$$

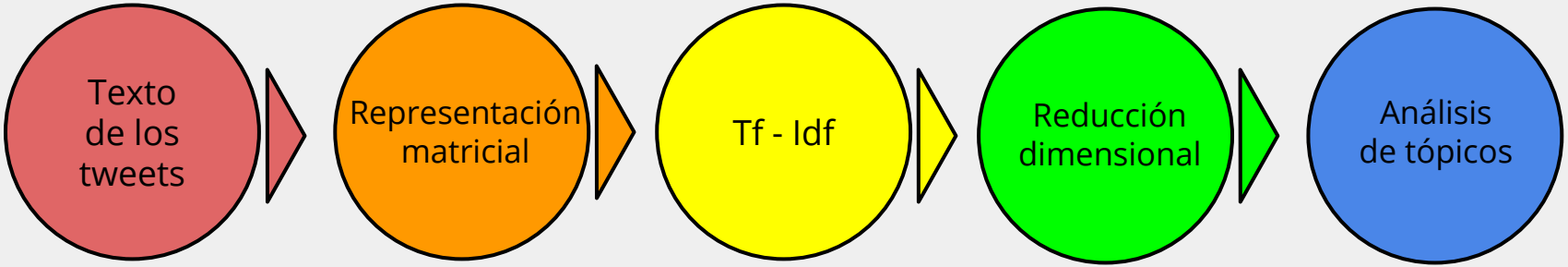
Donde H es la matriz de documentos por tópico y W una matriz de tópicos por término.

Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 267-273). ACM

Text mining


Texto de los tweets
Representación matricial
Tf - Idf
Reducción dimensional
Análisis de tópicos

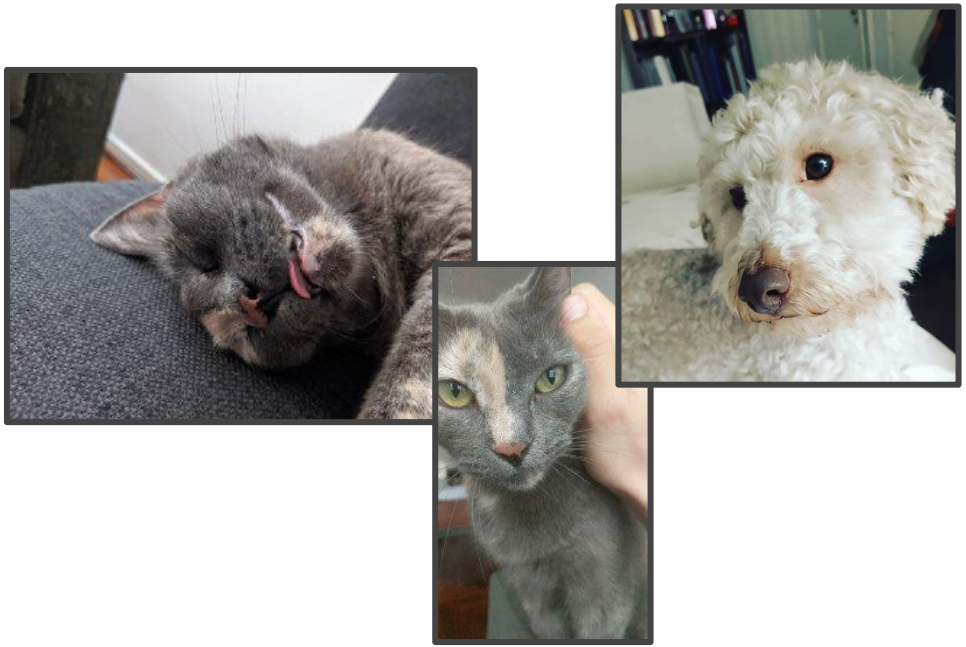
Text mining



- ***Santiago Maldonado:*** Desaparecido en agosto del 2017.
- ***Faltante de Boletas.***
- ***Autoridades de Mesa:*** Ausencia en las escuelas.
- ***Economía:*** centrado en pobreza, desempleo y programas de ajuste.
- ***Tragedia de "Once":*** En 2012 hubo un accidente de trenes donde murieron 51 personas.
- ***Encuestas.***
- ***Provincia de Santa Cruz:*** Una de las principales provincias opositores gobernada por Alicia Kirchner.
- ***Provincia de Buenos Aires:*** La provincia de mayor superficie y población de la Argentina
- ***Venezuela.***

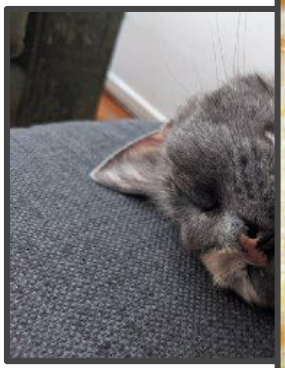
Modelo predictivo

Clasificación Supervisada



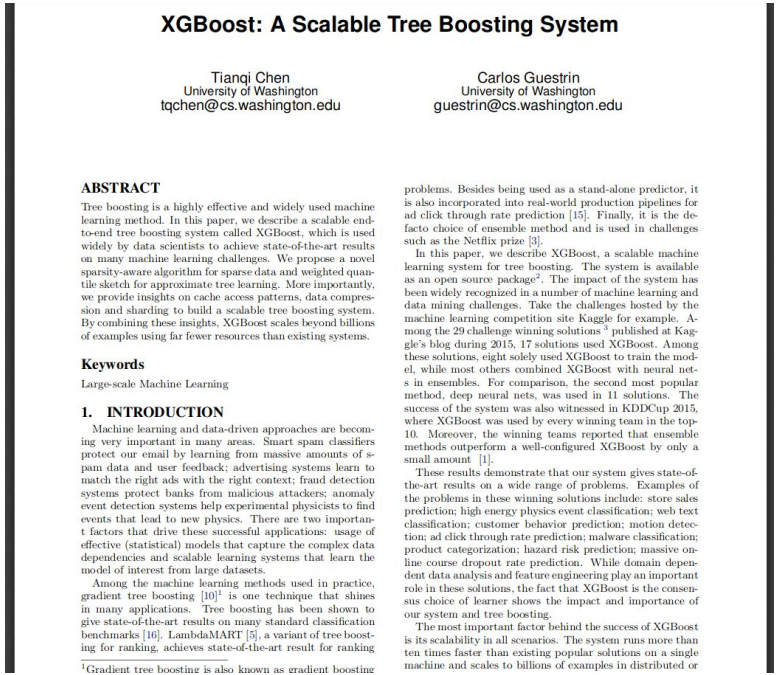
¿Gato o perro?

Clasific



perro?

XGBoost



XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

ABSTRACT

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

Keywords

Large-scale Machine Learning

1. INTRODUCTION

Machine learning and data-driven approaches are becoming very important in many areas. Smart spam classifiers protect our email by learning from massive amounts of spam data and user feedback; advertising systems learn to match the right ads with the right context; fraud detection systems protect banks from malicious attackers; anomaly event detection systems help experimental physicists to find events that lead to new physics. There are two important factors that drive these successful applications: usage of effective (statistical) models that capture the complex data dependencies and scalable learning systems that learn the model of interest from large datasets.

Among the machine learning methods used in practice, gradient tree boosting [10][1] is one technique that shines in many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks [16]. LambdaMART [5], a variant of tree boosting for ranking, achieves state-of-the-art result for ranking problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction [15]. Finally, it is the de-facto choice of ensemble method and is used in challenges such as the Netflix prize [3].

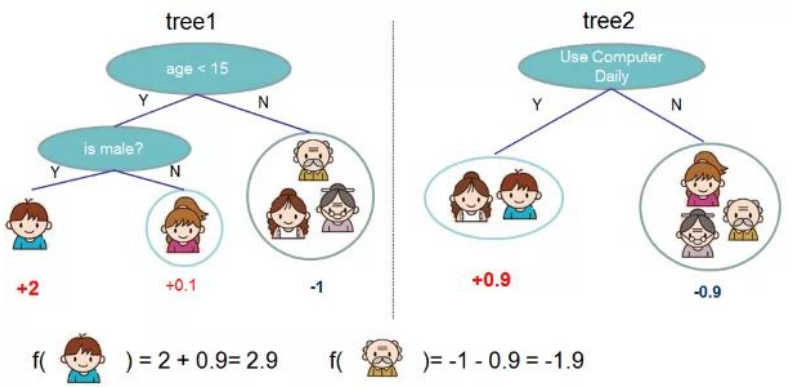
In this paper, we describe XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package[2]. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. Take the challenges hosted by the machine learning competition site Kaggle for example. Among the 29 challenge winning solutions [3] published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural nets in ensembles. For comparison, the second most popular method, deep neural nets, was used in 11 solutions. The success of the system was also witnessed in KDDCup 2015, where XGBoost was used by every winning team in the top-10. Moreover, the winning teams reported that ensemble methods outperform a well-configured XGBoost by only a small amount [1].

These results demonstrate that our system gives state-of-the-art results on a wide range of problems. Examples of the problems in these winning solutions include: store sales prediction; high energy physics event classification; web text classification; customer behavior prediction; motion detection; ad click through rate prediction; malware classification; product categorization; hazard risk prediction; massive online course dropout rate prediction. While domain dependent data analysis and feature engineering play an important role in these solutions, the fact that XGBoost is the consensus choice of learner shows the impact and importance of our system and tree boosting.

The most important factor behind the success of XGBoost is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or

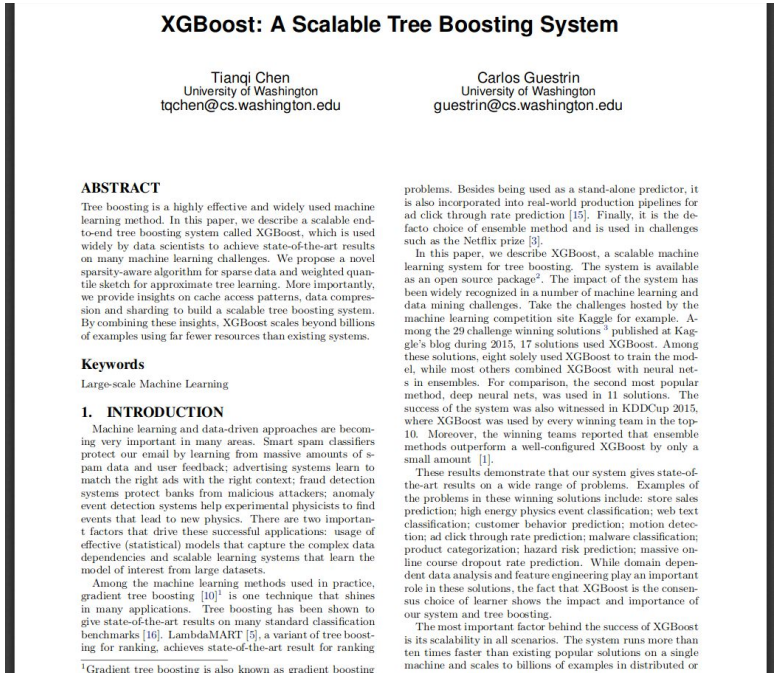
[1]Gradient tree boosting is also known as gradient boosting

Árboles de decisión



[1] Chen, T., & Guestrin, C. X. A scalable tree boosting system. CoRR. 2016. *arXiv preprint arXiv:1603.02754*.
[2] Nielsen, D. (2016). *Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?* (Master's thesis, NTNU).

XGBoost



XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

ABSTRACT

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

Keywords

Large-scale Machine Learning

1. INTRODUCTION

Machine learning and data-driven approaches are becoming very important in many areas. Smart spam classifiers protect our email by learning from massive amounts of spam data and user feedback; advertising systems learn to match the right ads with the right context; fraud detection systems protect banks from malicious attackers; anomaly event detection systems help experimental physicists to find events that lead to new physics. There are two important factors that drive these successful applications: usage of effective (statistical) models that capture the complex data dependencies and scalable learning systems that learn the model of interest from large datasets.

Among the machine learning methods used in practice, gradient tree boosting [10][1] is one technique that shines in many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks [16]. LambdaMART [5], a variant of tree boosting for ranking, achieves state-of-the-art result for ranking problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction [15]. Finally, it is the de-facto choice of ensemble method and is used in challenges such as the Netflix prize [3].

In this paper, we describe XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package[2]. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. Take the challenges hosted by the machine learning competition site Kaggle for example. Among the 29 challenge winning solutions [3] published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural nets in ensembles. For comparison, the second most popular method, deep neural nets, was used in 11 solutions. The success of the system was also witnessed in KDDCup 2015, where XGBoost was used by every winning team in the top-10. Moreover, the winning teams reported that ensemble methods outperform a well-configured XGBoost by only a small amount [1].

These results demonstrate that our system gives state-of-the-art results on a wide range of problems. Examples of the problems in these winning solutions include: store sales prediction; high energy physics event classification; web text classification; customer behavior prediction; motion detection; ad click through rate prediction; malware classification; product categorization; hazard risk prediction; massive online course dropout rate prediction. While domain dependent data analysis and feature engineering play an important role in these solutions, the fact that XGBoost is the consensus choice of learner shows the impact and importance of our system and tree boosting.

The most important factor behind the success of XGBoost is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or

[1]Gradient tree boosting is also known as gradient boosting

Pasos (forma simplificada):

1. Ajustar un modelo a los datos

   F1(X) = Y

2. Ajustar un modelo al gradiente de la loss del modelo anterior

   h1(X) = Y - F1(X)

3. Crear un nuevo modelo

   F2(X) = F1(X) + h1(X)

   donde F2 es la versión "boosted" de F1.

Y así sucesivamente

Fm(X) = F(m-1)(X) + y h(m-1)(X)

donde y es un factor de escala.

[1] Chen, T., & Guestrin, C. X. A scalable tree boosting system. CoRR. 2016. *arXiv preprint arXiv:1603.02754*.
[2] Nielsen, D. (2016). *Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?* (Master's thesis, NTNU).

Resultados

Modelo predictivo

Input:

Matriz donde cada fila es un usuario y donde cada columna es una feature del usuario.
Lista de features:

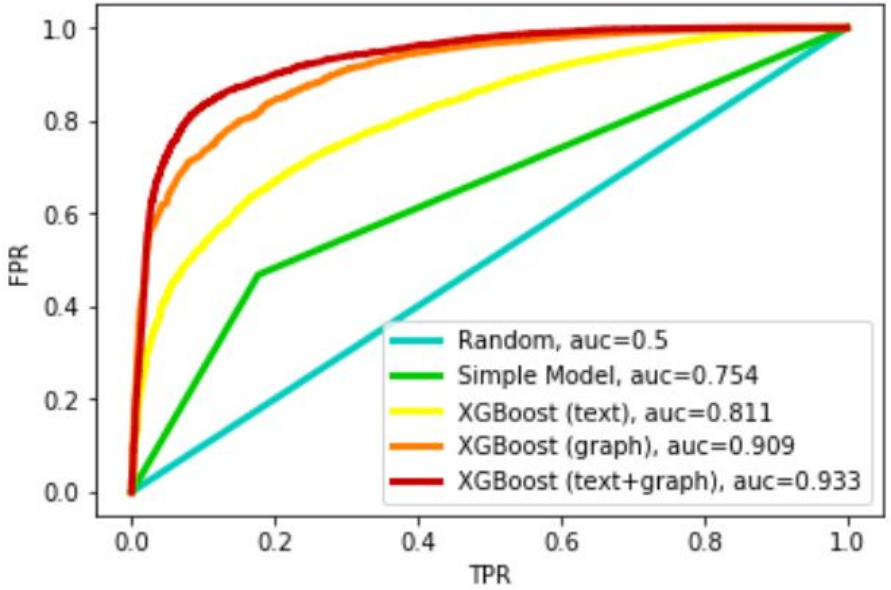
- ***Graph features:*** Degree, PageRank, Betweenness centrality, Clustering coefficient, Clustering affiliation (Louvain), z-score del grado interno, Coeficiente de participación.
- ***Text features:*** de qué tema habla.

Output:

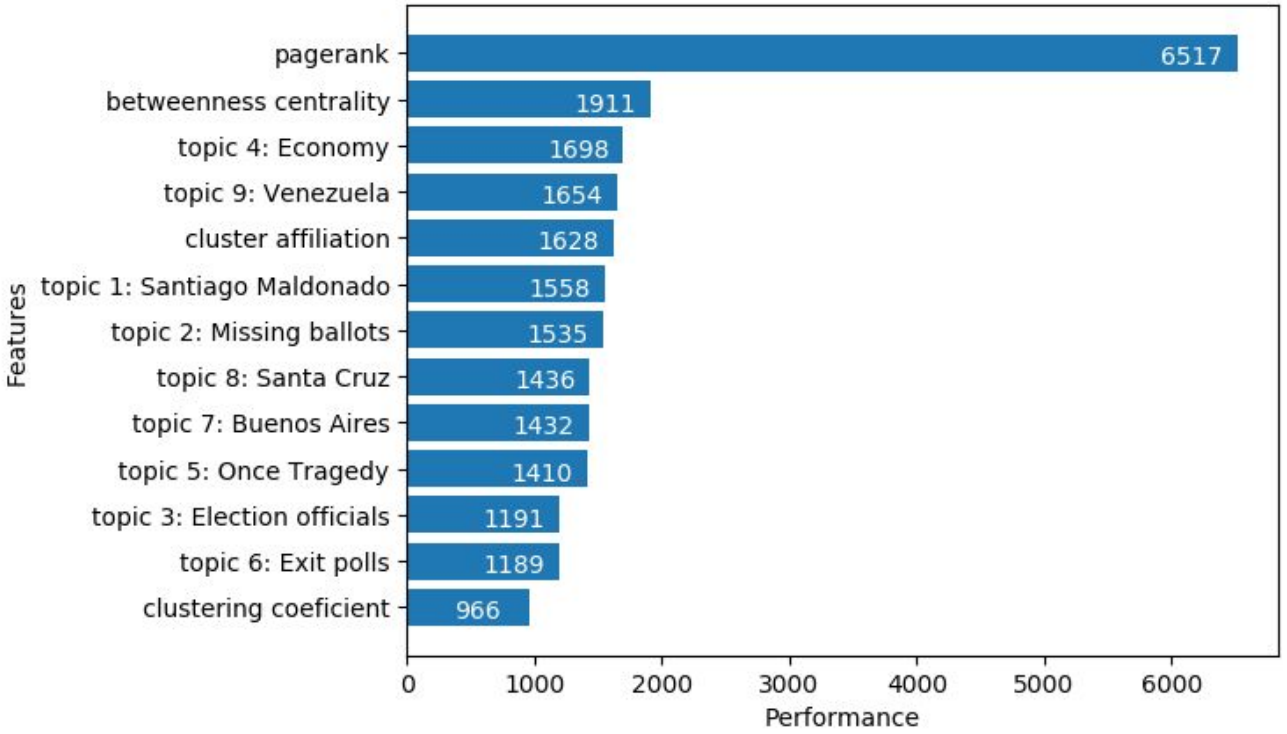
Vector de 1 y 0, cambió de comunidad entre elecciones o no respectivamente.

Evaluación del modelo

Curvas ROC

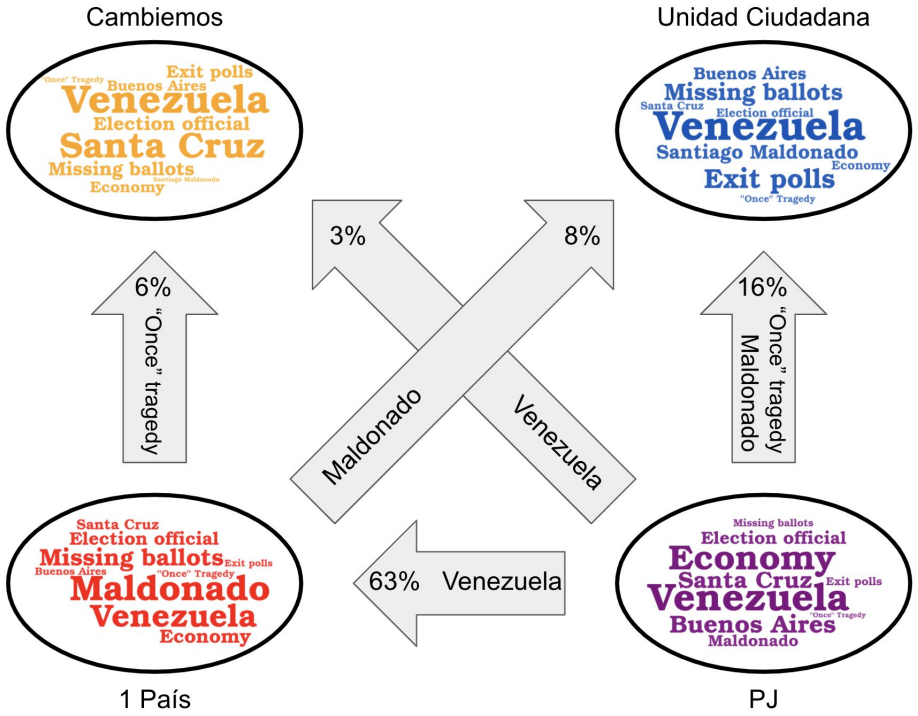


Resultados: Feature Importance



Tópicos más persuasivos

- ¿Quiénes cambian de grupo?
- ¿De qué habla la gente que cambia de grupo?
- ¿Y los que se quedan?



Resultados 2019

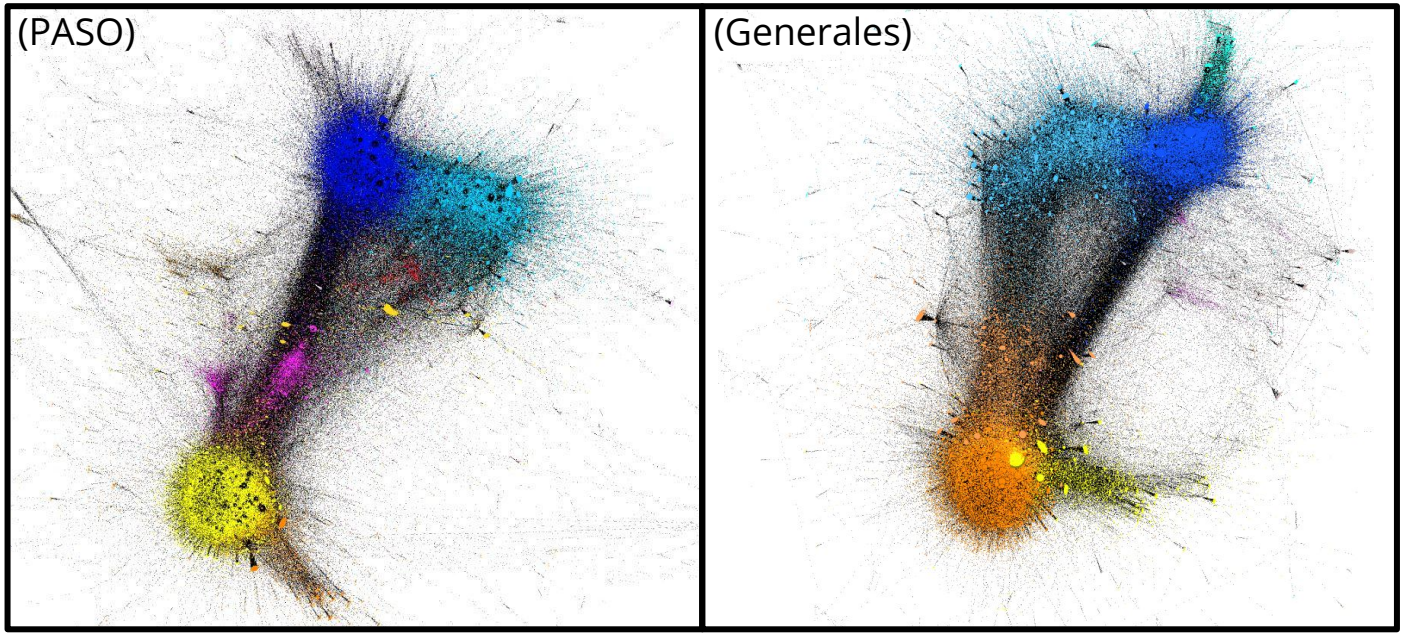
Datos 2019

Elecciones PASO 2019:

- 1.638.585 tweets
- 1.236.172 RT
- 256.860 usuarios

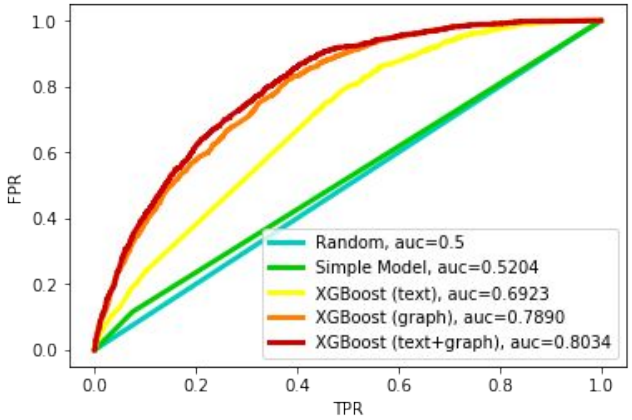
Elecciones Generales 2017:

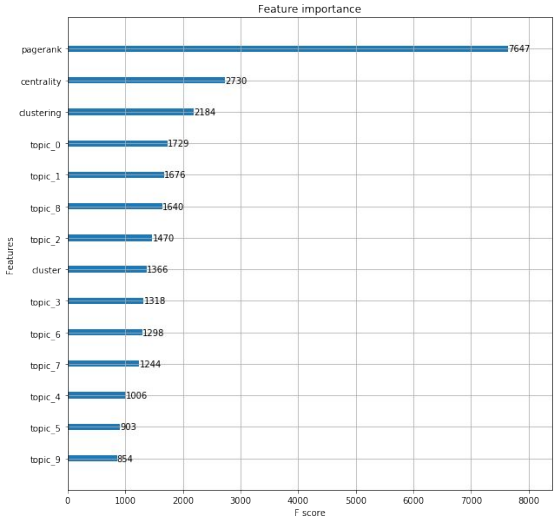
- 1.587.563 tweets
- 1.282.184 RT
- 271.910 usuarios



Resultados 2019

Curvas ROC





Tópicos más persuasivos:

Los de **Consenso Federal (Lavagna)** que se van a **Juntos por el Cambio (Macri)** hablan principalmente del "gobierno de la ciudad".

Mientras que los que se van al **Frente de Todos (Fernández)** hablan de Trabajo y Empleo (temas económicos).

Conclusiones

Conclusiones

Conclusiones

- El modelo predictivo, usando datos de twitter, efectivamente **detecta cambios en la afiliación política** de las personas (*auc= 0.933*) .

Conclusiones

- El modelo predictivo, usando datos de twitter, efectivamente **detecta cambios en la afiliación política** de las personas (*auc= 0.933*) .

- Mostramos la importancia de las **características topológicas** de los usuarios en la red de RT (comúnmente ignoradas en la literatura).

Conclusiones

- El modelo predictivo, usando datos de twitter, efectivamente **detecta cambios en la afiliación política** de las personas (*auc= 0.933*) .

- Mostramos la importancia de las **características topológicas** de los usuarios en la red de RT (comúnmente ignoradas en la literatura).

- Detectamos los **tópicos más persuasivos** e importantes. En una sociedad polarizada, esta herramienta tiene ***muchas aplicaciones*** para las ciencias sociales y para los partidos políticos.

Conclusiones

- El modelo predictivo, usando datos de twitter, efectivamente **detecta cambios en la afiliación política** de las personas (*auc= 0.933*) .

- Mostramos la importancia de las **características topológicas** de los usuarios en la red de RT (comúnmente ignoradas en la literatura).

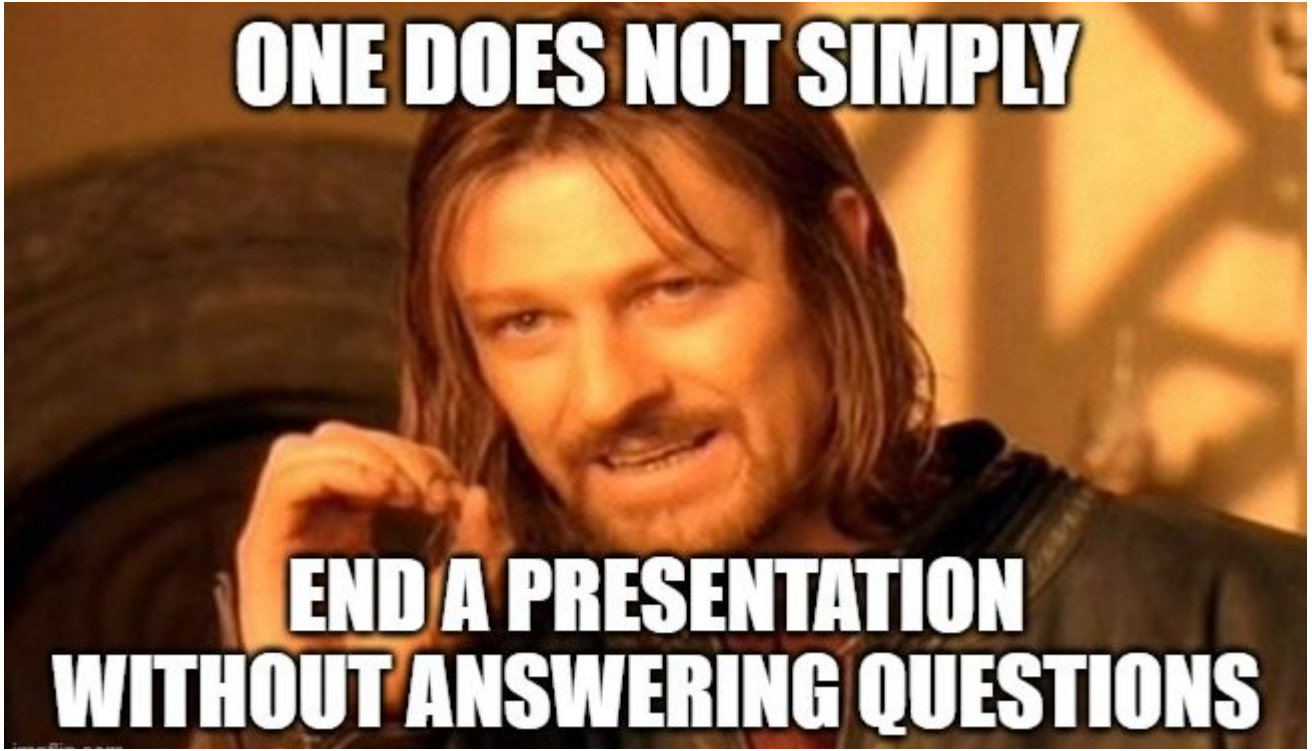
- Detectamos los **tópicos más persuasivos** e importantes. En una sociedad polarizada, esta herramienta tiene ***muchas aplicaciones*** para las ciencias sociales y para los partidos políticos.

- Los resultados son **consistentes** para distintas elecciones.

¡Muchas Gracias!

Federico Albanese
(falbanese@dc.uba.ar)
@f__albanese

Preguntas


ONE DOES NOT SIMPLY
END A PRESENTATION
WITHOUT ANSWERING QUESTIONS