Unveiling Community Dynamics on Instagram Political Network

Unveiling Community Dynamics on Instagram Political Network

Carlos H. G. Ferreira
Universidade Federal de Minas Gerais
Universidade Federal de Ouro Preto
Politecnico di Torino

Fabricio Murai, Ana P.C. Silva, Jussara Almeida
Universidade Federal de Minas Gerais

Martino Trevisan, Luca Vassio, Marco Mellia
Politecnico di Torino

Idilio Drago
University of Turin

ABSTRACT

Online Social Networks (OSNs) allow users to generate and consume content in an easy and personalized way. Among OSNs, Instagram has seen a surge in popularity, and political actors exploit it to reach people at scale, bypassing traditional media and often triggering harsh debates with and among followers. Uncovering the structural properties and dynamics of such interactions is paramount for understanding the online political debate. This is a challenging task due to both the size of the network and the nature of interactions.

In this paper, we define a probabilistic model to extract the backbone of the interaction network among Instagram commenters and, after that, we uncover communities. We apply our model to 10 weeks of comments centered around election times in Brazil and Italy. We monitor both politicians and other categories of influencers, finding persistent commenters, i.e., those who often comment together on Instagram posts.

Our methodology allows us to unveil interesting facts: i) commenters' networks are split into few communities; ii) community structure in politics is weaker than in general profiles, indicating that the political debate is a blur, with some commenters bridging strongly opposed political actors; and iii) communities engaging on political profiles are bigger, more active and more stable during electoral period.

KEYWORDS

Complex Networks; Dynamic User Modeling; Generative Models; Online Social Networks; Instagram; Politics

ACM Reference Format:
Carlos H. G. Ferreira, Fabricio Murai, Ana P.C. Silva, Jussara Almeida, Martino Trevisan, Luca Vassio, Marco Mellia, and Idilio Drago. 2020. Unveiling Community Dynamics on Instagram Political Network. In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394231.3397913

1 INTRODUCTION

Online Social Networks (OSNs) have become a predominant means for direct communication, allowing users to generate and consume content in an easy, fast and personalized way. OSNs are nowadays a core component for information dissemination, marketing and advertisement [14, 26]. Instagram[1] has recently seen a growth in popularity, for its dynamic and user-centric interface, which is populated with rich multimedia content. A whole ecosystem of *influencers* has emerged in the platform, i.e., users who gain popularity by posting influential content or by leading the debate when commenting on others' posts [11].

Instagram is now a relevant channel to influence society. As such, also politicians and political personalities exploit it to reach the population at scale. Whereas the role of OSNs on social mobilization and political engagement is well studied, few efforts focus on how these interactions occur via Instagram. Previous works analyze user engagement based on content type [24], the impact of the posted content in marketing contexts [10], users' interests in political content [14] and the characteristics of comments left on political content [21]. However, the structural properties of the networks formed from user interactions are not understood yet. Revealing how interactions take place on Instagram and in the political context, is thus paramount for understanding the impact of the online debate in our society.

In Instagram jargon, a *profile* is followed by a set of people – i.e., its *followers*. A profile with a large number of followers is called an *influencer*. Influencers post content (i.e., a *post*). The profile's followers – and anyone registered on Instagram in the case of public profiles – can view and put a like on posts, and possibly comment on them, becoming *commenters*. Here, we are interested in commenters that comment on the same post, i.e., *co-commenters*. The intuition is that two profiles may be engaged in an online debate, hinting to structures that may be natural (i.e., based on interests) or driven by hidden patterns (e.g., ad-campaign or coordinated behavior).

The Instagram ecosystem features high-order structures, characterized by a superposition of many highly connected graphs, which are the structures that emerge from users who comment on the same post (co-commenters). Finding and observing communities based on co-commenters over time is challenging. First, two commenters may become co-commenters by chance. Some pairs can suddenly interact and never co-comment again. Other patterns can consistently repeat over time [1, 2]. Co-commenters occur organically because the set of users following each profile varies. Even for a single influencer, its commenters may be interested in specific topics. Commenters' interests also change over time, e.g., commenting solely when a debate heats up. Some posts are extremely popular, e.g., because of their controversial topics, thus attracting the attention of a large share of commenters. These characteristics make it difficult to distinguish structural patterns from sporadic, weak or incidental interactions [6, 15].

To address these challenges, we here model commenters' activity on Instagram as networks where posts create interaction opportunities. We consider all posts from homogeneous groups of influencers, e.g., athletes, singers, or politicians, that have been posted in the

[1] https://www.instagram.com/

same week. Given such a snapshot, we create a projected network where nodes represent commenters. There is an edge between any two commenters if both have commented on the same post, i.e., they are co-commenters. The edge weight accounts for the number of posts in which they both commented. We then define a probabilistic network model where the edge weight distribution in the projected network depends on commenters' activity levels and the number of opportunities they have to interact. Contrasting the Instagram data with this probabilistic model (called null model), we extract the *backbone* of the network by pruning edges whose weights are within the expected range under independent commenters' behavior assumption. Intuitively, we retain only those co-commenters that interacted more often than one would expect if interactions would happen independently. We then extract communities from the backbone network using the Louvain algorithm [3] and characterize the obtained communities.

As case studies, we analyze a large dataset of Instagram comments containing the activity of approximately 3 *million* commenters on 36 824 posts of 320 influencers. These influencers include profiles of popular political figures – 80 from Italy and 80 from Brazil, as well as 160 top-influencers in other categories (e.g., sportsmen, celebrities, musicians), 80 from each country, which we use as a baseline. We focus on two months surrounding elections that took place in each country. We track all posts of each influencer in the period, recording all comments in those posts.

Our goal is to study co-commenters' networks to understand how people interact on the Instagram ecosystem. Our research questions (RQ) and respective findings are:

- **RQ1: What structures emerge from the backbone network of co-commenters on political discussions? Are these structures significantly different from discussions on other topics?**
  In the backbone networks, we identify fewer and better-defined communities than in the original networks. In politics, these communities are weaker than in general topics, indicating that some commenters act as bridges among political actors. This holds true in both countries.
- **RQ2: How these structures evolve over time?**
  By observing 10 weeks around political elections, we see the heating up and sudden disinterest in the online debate before and after elections. Unlike in general topics where communities are persistent and consistent over time, in the political context we observe that the fraction of engaged commenters grows and plumbs. Communities change substantially over time, with most commenters with the highest activity levels that remain consistently active over time.

In summary, we are the first to provide a large-scale study of Instagram interactions of co-commenters, analyzing millions of commenters in distinct cultural perspectives. We provide both a methodology to highlight common trends and particularities as well as a characterization of communities of co-commenters with a high level of engagement and evidence of coordination. To foster further studies, the backbone networks extracted by our methodology is available to researchers upon request.

This paper is organized as follows. Section 2 summarizes related work. Section 3 describes our dataset. Section 4 presents our methodology to extract backbone networks and finding communities of co-commenters. Section 5 analyses the structures and communities emerging from the backbone networks, while Section 6 focuses on the community dynamics. Finally, Section 7 concludes the paper.

2 RELATED WORK

We are not the first to study users' behavior on OSNs in the political context. Authors of [7] study users' interactions, highlighting provoking and humorous posts in the political debate via Twitter. Authors of [20] analyze the political debate around *Brexit* via Facebook. Both works illustrate how users' behavior is strongly associated with intentions, e.g., the dissemination of ideas or misinformation through users' interactions. More recently, Whatsapp has been explored in the political debate. Authors of [18] show that Whatsapp is a source of political debate and propagation of fake news and misinformation. In contrast to those works, we here focus on the dynamics of the graph formed around co-commenters on Instagram.

Few efforts have been made to understand how users interact on Instagram, especially considering dynamic behaviors. In [24], authors analyze users' engagement of 12 Instagram profiles divided into classes (i.e., sports, news and politics). Authors search for *impersonators*, those who simulate others' behaviors to perform specific activities, such as spreading fake news. The work in [10] analyzes the influence of content posted for social media marketing. Similarly, authors of [14] study images posted by candidates during the US 2016 primary elections, highlighting combined factors that attract user engagement on posts.

We have performed a quantitative study of the political debate via Instagram in previous work [21]. Our results show that the profiles of politicians have significantly more interactions than others. Users comment more and for longer periods in politics, with a large number of replies that are usually not explicitly requested. In contrast, we here focus on networks of co-commenters, studying their structure and dynamic behaviors. We are the first to evaluate such aspects in two countries and multiple scenarios, analyzing a large number of users in a long time span.

We propose a methodology to extract backbones from co-commenters networks that reveals relevant connections based on substantiated statistical criteria. Some works that follow a similar approach in other domain-specific scenarios include: recommendation systems under online commercial networks [25], urban mobility patterns [13], transportation network [23], congressional co-voting network [5, 6] and citations networks [19].

3 DATASET

In this section we describe our dataset starting from its collection process and preprocessing. Then, we provide an overview and discussion of the dataset.

3.1 Crawling

We collect data from Instagram profiles in Brazil and Italy. We focus on electoral periods to capture the political debates taking place on the social network. In Brazil, we focus on Instagram posts

Table 1: Dataset characteristics for each of the 10 weeks. In bold, the weeks of the elections in the respective country.

Week	Politics				General			
	Brazil		Italy		Brazil		Italy	
	# Posts	# Commenters	# Posts	# Commenters	# Posts	# Commenters	# Posts	# Commenters
1	1 487	37 406	779	17 427	746	172 454	733	54 407
2	1 648	67 799	739	20 873	778	180 711	703	49 290
3	1 798	103 506	742	20 876	719	164 040	594	52 052
4	1 951	94 327	907	21 402	854	186 333	649	54 677
5	**2 307**	**145 618**	1 080	22 029	680	125 414	683	52 318
6	958	184 993	1 240	22 890	771	158 522	720	69 066
7	1 195	123 797	**1 316**	**26 600**	723	131 563	657	61 168
8	**1 400**	**145 499**	701	31 308	798	152 705	635	66 337
9	799	191 282	762	17 171	733	146 128	540	31 520
10	606	50 546	656	19 926	763	159 628	507	33 781

published during the national general elections of October 7th (first round) and October 28th (second round), 2018. We include weeks before and after the election dates, monitoring posts of selected profiles from September 2nd until November 10th, 2018. Similarly, in Italy we observe the European elections held on May 26th, 2019, collecting data about posts published from April 7th to June 15th.

We use a custom web crawler to scrape data from Instagram that relies on the Instaloader library[2]. We perform the crawling in September 2019, downloading only posts seen in the electoral periods mentioned above. Given a profile, the crawler looks for posts and downloads metadata and all comments these posts received from any Instagram user, including comments received *after* the electoral periods. Note that interest in posts decreases sharply with time [21]. We focus only on *public* Instagram profiles and posts, collecting all visible comments they receive. We perform the crawler respecting Instagram policies to avoid overloading the service. Moreover, we avoid collecting any sensitive information of commenters, such as display name, photos, or any other metatada, even if public.

We select two groups of influencers to follow:

- *Politics*: we list the most popular Brazilian and Italian politicians and official political parties profiles. We consider 80 profiles who have a verified Instagram account and a sizable number of posts for Brazil and another 80 for Italy. In total, they posted 14 149 and 8 922 posts that received more than 8 million and 1.9 million comments, respectively.
- *General*: we collect posts for non-political influencers that we use as a control group. We rely on HypeAuditor[3] rank to obtain the list of most popular profiles for the Sport, Music, Show, and Cooking categories. Similar to the *Politics* groups, we pick 80 profiles for Brazil (Italy) that created 7 565 (6 421) posts and received 15 million (14 million) comments.

3.2 Data preprocessing

To build the network of co-commenters, we consider all posts published in each week for each country and category of influencers.[4] We consider a week-long interval as a natural trade-off to aggregate enough posts. We then observe all comments these posts received to build the co-commenters' network. As we have 10 weeks of data, we build 10 networks for each country and category. We call each of such 40 networks a weekly-snapshot, or a *week* for simplicity.

To reduce noise introduced by occasional commenters we remove those commenting in just one post in a given snapshot. Intuitively, those leaving a single comment can hardly belong to significant communities of co-commenters. As such these commenters do not have an impact on the backbone networks we build. This step removes 70–85% of the commenters, depending on the dataset and the considered week. We observe that 95% of the removed commenters commented less than three times considering all weeks and categories. Given their low activity, only 30–55% of the comments are ignored. All results in the following refer to the dataset after the removal of these occasional commenters.

3.3 Dataset characterization

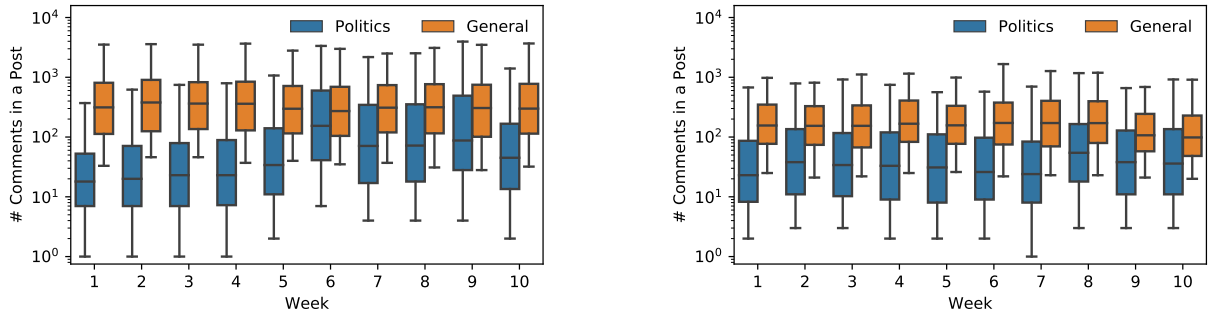
We present a high-level overview of each snapshot in Table 1. It details the number of posts and commenters for Politics and General influencers in Brazil and Italy. Elections (in bold) took place on the Sunday of the 5th and 8th weeks in Brazil, and on the Sunday of 7th week in Italy. Considering politics, we observe the largest number of posts on the week of the elections, with a steady increase in the number of posts that plumbs immediately after elections. Interestingly, the largest number of commenters appears on the week immediately after the elections – manual inspection suggests this effect is due to celebrations by supporters. If we consider the General category, the number of posts and commenters is rather constant, with a slight decrease in the last two weeks for Italy, which approach the summer holidays.

We complement the analysis with the distribution of the number of comments each post received for each week (Figure 1). We use box-plots for ease of visualization. The black stroke represents the median. Boxes span from the 1st to the 3rd quartiles, whiskers mark the 5th and the 95th percentiles. Each post from politics receives in median few tens of comments, while posts from non-politicians 10 times as much (notice the log y-axis). This is explained by the large number of celebrities and public figures (e.g., singers, actors, athletes) that attract lots of comments per commenter (see Table 1). These considerations hold for both countries. Considering evolution over time, we note peaks on the weeks around the elections and immediately after, when the number of comments per post increases by an order of magnitude for Brazil.

4 METHODOLOGY

We now formalize the network model of co-commenters on Instagram. Then, we propose a null model that guides the identification of non-casual interactions, which will be part of backbone networks.

[2]https://instaloader.github.io
[3]https://hypeauditor.com/
[4]We consider weeks starting on Monday and ending on Sunday.



(a) Brazil. **(b) Italy.**

Figure 1: Number of comments per post in different weeks for the two categories and countries. Notice the log scale in y-axis.

Finally, we illustrate how we apply the Louvain algorithm to detect communities in those backbone networks.

4.1 Network modeling

We model the temporal dynamics of commenters in each weekly-snapshot using graphs. Given a week w, we take the corresponding set of posts P to create a weighted and undirected graph $G_P = (V_P, E_P)$. We have a different P, and hence a different graph G_P, for each week, country and category. The set of vertices V_P represents commenters that in week w commented in at least two posts in P. E_P is the set of undirected edges, where there exists an edge e_{cd}=(c, d) if commenters c and d commented in at least one common post in P. That is, edges link co-commenters. The weight $\gamma(cd) \in \mathbb{N}^+$ of edge e_{cd} is defined as the number of posts where the co-commenters commented together, i.e., $\gamma(cd) \in \{1, 2, \ldots, |P|\}$.

Each post p in P will generate a clique, i.e., a graph composed by a subset of vertices (commenters) such that every two distinct vertices in the clique are adjacent. Hence, each subgraph induced by a single post p is complete by definition. The final network G_P we obtain can be seen as the superposition of all the cliques generated by the posts in P.

The network G_P represents all interactions that happened among co-commenters on posts of a group of influencers during a week. G_P results in a complex network with many nodes and a very large number of edges. In the following section, we show how to extract a more tractable and informative subgraph of G_P that contains only its most important edges.

4.2 Extracting the network backbone

Complex networks are non-random in many aspects. An important question is how to quantify the statistical significance of an observed network structure with respect to a given random graph model [4]. Null models are crucial in determining whether networks display certain features to a greater extent than expected by chance under a null hypothesis. A null model matches some of the features of a graph and satisfies a collection of constraints, but is otherwise taken to be an unbiased random structure. It is used as a baseline to verify whether the object in question displays some non-trivial features, i.e., properties that would not be expected on the basis of chance alone or as a consequence of the constraints. An appropriate null model behaves in accordance with a reasonable null hypothesis for the behavior of the system under investigation. Rather than randomizing an existing network, null models can be constructed by generative growing networks.

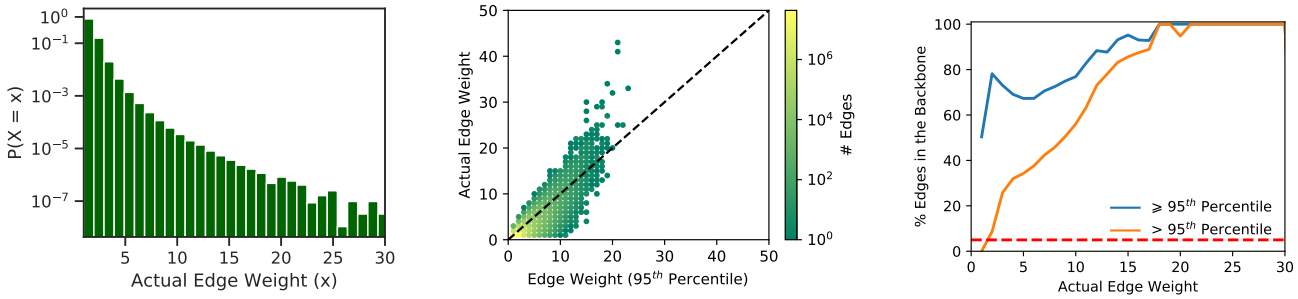
Our goal is to create a null model $\hat{G}_P$ from our network G_P, provided the hypothesis that the commenters behave independently. Then, we can observe the edges of the real network that do not behave in accordance with the null model. Such edges will compose the backbone of the network. The idea of the *backbone network* is to capture only the *salient* interactions, which are those that strongly suggest non-random interactions. Intuitively, we want to highlight those co-commenters that appear to co-comment more frequently than expected if they would behave independently.

We build a null model $\hat{G}_P$ in which edge weights are defined under a generative process in which commenters act independently of each other. However, their interactions with influencers' posts are not identically distributed. Our generative model takes into account the the popularity of a post and the engagement of each commenter for a given influencer. In other words, comments are randomly assigned to commenters preserving: i) the set of influencers on which each commenter writes a comment; ii) the popularity of a specific post (number of unique commenters) and iii) the relative ratio of comments per commenter on a given influencer's posts.

Let I be the set of all influencers who wrote posts in P. Given a week w, let $C_p \subseteq V_P$ be the set of unique commenters in post $p \in P$ and $\{\mathcal{P}_i\}_{i \in I}$ be a partitioning of P based on the influencer $i \in \mathcal{I}$ who created the post. The total number of posts in $\mathcal{P}_i$ commented by c is $x_i(c) = \sum_{p \in \mathcal{P}_i} \mathbb{1}\{c \in C_p\}$, where $\mathbb{1}\{.\}$ is the identity function. We define c's *engagement* relative to other commenters in i's posts as $f_i(c) = x_i(c)/\sum_d x_i(d)$. Now we can describe in details the three steps of the generative model to create our null model $\hat{G}_P$:

1) For each post $p \in P$, we consider a random assignment of each of the $|C_p|$ comments to a commenter $c \in V_P$ with probability $f_i(c)$, where i is the author of p. Hence, the probability that c is assigned to at least one comment in $p \in \mathcal{P}_i$ is $r_p(c) = 1 - (1 - f_i(c))^{|C_p|}$.

2) For each pair of commenters c and d, we denote by $q_p(cd)$ the probability that both get assigned to post p and by $r_p(d|c)$ the probability that d gets assigned to p conditioned on c getting assigned to p. We approximate $q_p(cd) = r_p(c)r_p(d|c) \approx r_p(c)r_p(d)$, for each $p \in P$. This approximation works well when $|C_p|$ is large,



(a) Probability distribution of edge weights $\gamma(cd)$ in the original graph G_P.

(b) Actual edge weight $\gamma(cd)$ compared with null model edge weight $\hat{\gamma}_{95}(cd)$.

(c) Fraction of edges included in the backbone B_P, separately for different values of γ.

Figure 2: Network characteristics for posts of influencers for Brazil - Politics (Week 1).

as in the case of influencers' posts. Then, for each post $p \in P$ we build a clique among all possible commenters in V_P, where the edge weight between commenters c and d is a random variable $\hat{\Gamma}_p(cd) = \text{Bernoulli}(q_p(cd))$.

3) The null model $\hat{G}_P = (\hat{V}_P, \hat{E}_P)$ is composed by the superposition of all the created cliques. Hence, an edge $\hat{e}_{cd} \in \hat{E}_P$ will have a weight distribution described by a random variable $\hat{\Gamma}(cd) = \sum_{p \in P} \hat{\Gamma}_p(cd)$. Therefore, it will be a sum of Bernoulli random variables with distinct probabilities [22]. The distribution of $\hat{\Gamma}(cd)$ is a Poisson Binomial distribution with parameters $q_1(cd), q_2(cd), \ldots, q_{|P|}(cd)$.

Having the null model $\hat{G}_P$, we can compare it with the actually observed network G_P. Comparing the two graphs, we create the backbone graph B_P of the original G_P networks by considering only edges in G_P whose weights have values exceeding by a large margin the ones expected in $\hat{G}_P$. Specifically, for each edge $\hat{e}_{cd}$ we compute the 95th percentile $\hat{\gamma}_{95}(cd)$ of its edge weight distribution $\hat{\Gamma}(cd)$. Then we compare this value $\hat{\gamma}_{95}(cd)$ with $\gamma(cd)$. If $\gamma(cd) > \hat{\gamma}_{95}(cd)$, we keep this edge in the backbone network. Intuitively, we keep only edges between co-commenters that co-commented in so many posts that, under the independence assumption, would be observed in 5% of cases only.

The 95th percentiles are computed independently for each edge from the random variable $\hat{\Gamma}(cd)$. For such a Poisson binomial distribution, there is a closed form for computing the 95th percentile and it relies on a combinatorial computation [9]. Since this is expensive to compute exactly, we use the Refined Normal Approximation (RNA) [9], a method that proved very good performance with low computational complexity.

4.3 Detecting communities

From the edges of the backbone network B_P, we extract communities using the Louvain algorithm [3, 16], widely used for community detection in networks [8, 17]. Communities are non-overlapping sets of nodes in the graph. The goal of Louvain algorithm is to maximize the modularity of the communities. Modularity is a measure between -1/2 and +1 that captures how much higher/lower is the density of edges inside communities in comparison to a network with the same degree sequence but where nodes are randomly connected. Modularity of 0.5 or higher is considered a solid indication of well-shaped communities. Since trying out all possible communities is computationally impractical, heuristic algorithms have been proposed. In Louvain algorithm, small communities are found first by optimizing modularity locally on all nodes. Then, each small community is grouped into one metanode and the first step is repeated. We refer readers to [3] for details.

The number of communities is not a parameter, but a result of the optimization procedure. Hence, given a network, we obtain the number of communities that provides the highest modularity value.

5 RQ1: COMMUNITY STRUCTURES

We now describe the community structure emerging from our data. We first show how our methodology extracts salient interactions from the original graph. Then, we characterize the communities and highlight insights emerging from co-commenters backbones.

5.1 The network backbones

We first show how our methodology builds the network backbones B_P. We characterize the original graph of interactions G_P and show how our model chooses the subset of edges whose weights deviate from the expected values. We use as running example the 1st week snapshot of the Brazilian Politics scenario.

Figure 2a shows the normalized distribution of the edge weights in the original graph. The normalization is performed such that the count in a histogram class is divided by the total number of observations in all classes. Notice that 82% of edges have weight equal to 1, i.e., the majority of co-commenters co-comment in a single post. Higher weights are less frequent (notice the log scale on the y-axis). Yet, some co-commenters share more than 20 posts.

The question then arises: *Are these weights expected?* We want to find those edges whose weight differs significantly from the independent behavior assumption. The scatter plot in Figure 2b compares the actual weight in G_P and the 95th percentile of expected weight in $\hat{G}_P$ for the same edge. Colors represent the number of edges (lighter colors represent larger quantities). First, most edges

Table 2: Characteristics of the original network G_P and backbone network B_P for Brazil - Politics (Week 1).

Network	# Nodes	# Edges	# Comm	Modularity
Original	37 394	74 095 748	6	0.22
Backbone	26 442 (70.7 %)	1 060 782 (1.4 %)	19	0.59

Table 3: Breakdown of backbone and communities over different weeks for Brazil - Politics. In bold, the weeks of the elections.

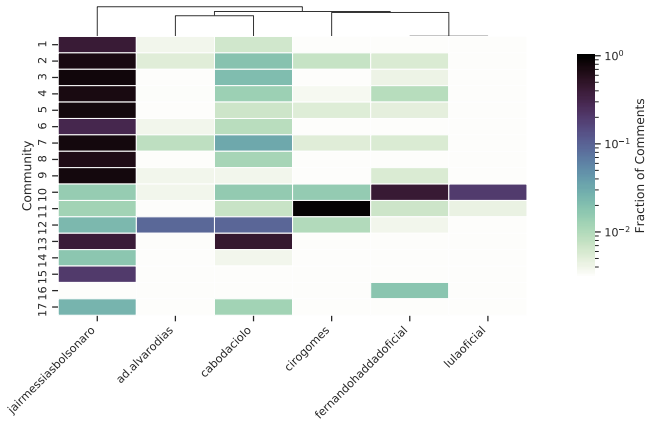
Week	% Nodes	% Edges	% Edges $\gamma(cd) > 1$	# Comm	Mod.
1	70.70	1.40	11.43	19	0.59
2	93.36	2.11	12.19	27	0.64
3	73.81	1.01	4.75	20	0.52
4	93.63	2.23	15.10	32	0.69
5	**94.30**	**2.65**	**19.36**	**17**	**0.61**
6	91.49	2.36	19.37	31	0.66
7	94.05	1.87	15.45	31	0.66
8	**95.40**	**2.13**	**15.29**	**27**	**0.64**
9	68.01	0.62	4.06	24	0.59
10	71.33	1.11	7.21	29	0.61

have a very low actual and estimated weights – notice the lightest colors for weights 1 and 2 in the bottom left corner. Second, edges above the main diagonal are those of interest for us, i.e., those that exceed the 95th percentile of the expected weight. Interestingly, the fraction of edges over the diagonal is higher for larger values of weights. This behavior indicates that co-commenters interacting on many posts deviate from the expected more often than under independent assumption.

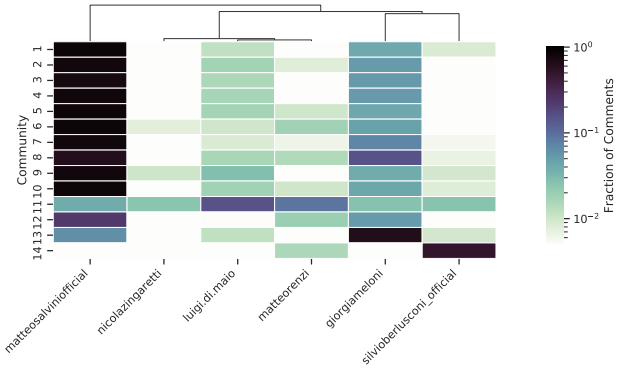
To dig into this aspect, we show in Figure 2c the percentage of edges that are included in the network backbone for each observed edge weight. If the null model holds true, we expect 5% of the edges to be included (those exceeding the 95th percentile) – highlighted by the red dotted line. Clearly, in the actual graph G_P edge weights do not follow the null hypothesis. The higher the weight is, the larger is the probability that co-occurrence exceeds the threshold.

However, G_P edge weights are integer numbers, and our generative model provides discrete distributions. The computation of percentiles is thus critical since the same value can refer to a *range* of percentiles. We have a rounding issue that is critical for low values. Filtering weights *greater than* or *greater or equal to* particular values results in significant differences for low weights. Figure 2c illustrates it by reporting the fraction of edges that would be included in the backbone in the two cases. Using *greater than* corresponds to a conservative choice since we include only edges for which the expected weight is strictly higher than the 95th percentile (orange curve). Notice how the number of edges in the backbone is reduced in this case, in particular for low weights. Conversely, *greater or equal to* would preserve more edges, including those whose weight corresponds possibly to a lower percentile (blue curve). We here stick to a conservative choice and keep edges whose actual weight is strictly greater than the 95th percentile.

Table 2 summarizes the resulting backbone network. Our approach discards 98.6 % of the edges – i.e., the vast majority of them is not salient and, thus, not included in the backbone. We then remove 29% of nodes, which are the ones for which no edge remains. To



(a) Brazil (1st round)



(b) Italy

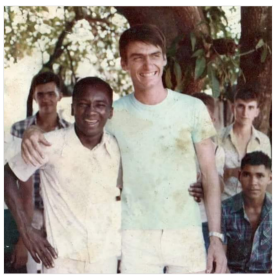
Figure 3: Distribution of comments among political leaders and communities during the main election weeks.

highlight the benefits of our approach, we show the number of communities and their modularity in the original and backbone graphs. In a nutshell, the Louvain algorithm fails to identify communities in the original graph - showing just 6 communities with quite poor modularity. After pruning uninteresting edges, we can identify a larger number of much better-shaped communities (modularity passes from 0.22 to 0.59). This result is particularly interesting as it shows the ability of our approach in finding cohesive groups of co-commenters. We provide a deep discussion on the community structure in the sections that follow.

Table 3 summarizes the main characteristics of the backbone networks obtained on each week of the Brazil - Politics scenario. Focusing on the first four columns, notice that we still include the majority of nodes, with percentages ranging from 68% to 95%. Considering edges, the percentage is always low (0.6–2.6%). The fourth column reports the fraction on edges in the backbone having weight larger than 1. Remind that, by design, a random behavior would lead to 5% of edges in the backbone, while here we observe up to 19%, despite our conservative filtering criteria. That is, by construction, we focus on those commenters with strong coordination. These results are stable and consistent over time.

(a) Community 3 - Post about a rally in São Paulo. www.instagram.com/p/BoXpvV6Hrkk

(b) Community 3 - Post about a rally in Vitória. www.instagram.com/p/BoXMwvwn6xj



(c) Community 7 - Post discussing racism. www.instagram.com/p/BomRItfH9p8

(d) Community 7 - Another post discussing racism. www.instagram.com/p/Boe7fQcHfJB

Figure 4: Examples of posts by Jair Bolsonaro (jairmessiasbolsonaro) in which two communities were more active.

Table 4: Networks backbone and identified communities for Brazil (BR) and Italy (IT).

Scenario	% Nodes	% Edges	% Edges $\gamma(cd) > 1$	# Comm	Mod
BR Politics	84.61	1.81	12.42	26	0.62
IT Politics	87.33	3.39	21.79	11	0.44
BR General	65.35	0.82	8.83	81	0.79
IT General	60.03	2.23	12.57	48	0.72

5.2 Communities of commenters

We now study the communities that we obtain from backbone graphs. The last two columns of Table 3 (Brazil - Politics) show that we obtain from 19 to 32 communities, depending on the week. Modularity values are rather high (always above 0.5), meaning that the community structure is strong.

We summarize results for other scenarios in Table 4, reporting average values across the 10 weeks for each scenario. Focus on Politics first and compare Brazil and Italy (first two rows). We observe similar percentages of nodes in the backbone networks. For Italy a larger fraction of edges are retained, potentially because of the smaller volume of profiles and comments (see Section 3). In both cases, we obtain a reasonable number of communities, a bit larger in Brazil with higher values of modularity than in Italy. For comparison - modularity is much lower if we consider the original graphs, and the number of communities would also be much lower. This result confirms the benefits already shown in Table 2.

Moving to the General scenarios (3rd and 4th rows), we notice that fewer nodes and edges are in the backbones when compared to the Politics backbones. More interestingly, we identify more and stronger communities. We root this phenomenon in the natural heterogeneity of the General scenarios that include influencers with different focuses, potentially attracting profiles with different interests. Not shown here for the sake of brevity, we observe that the communities strongly group commenters with different interests – e.g., we find communities interested in Sports, others focusing on Music, etc. Those communities are naturally more separated.

The scenario is more tangled when it comes to politics. Even if we find rather strong communities, some of them are *across-the-board* and include profiles commenting on politicians of different parties and embracing different topics. We provide a thorough analysis of communities in the Politics scenario in the next section.

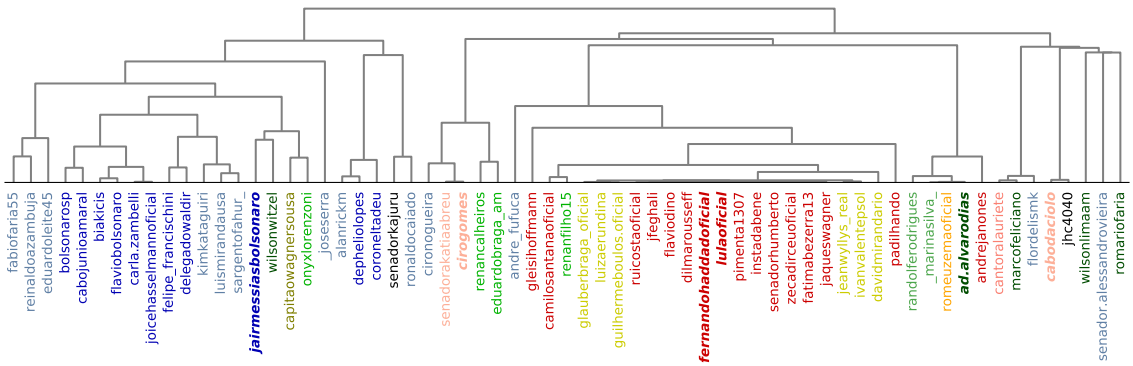
5.3 Analysis of political communities

We now investigate the interests of the communities for the Politics scenarios and show how the activity of commenters spreads across political profiles of different parties. Here, we focus on the election weeks for both countries to better capture the peak of the political debate on Instagram.

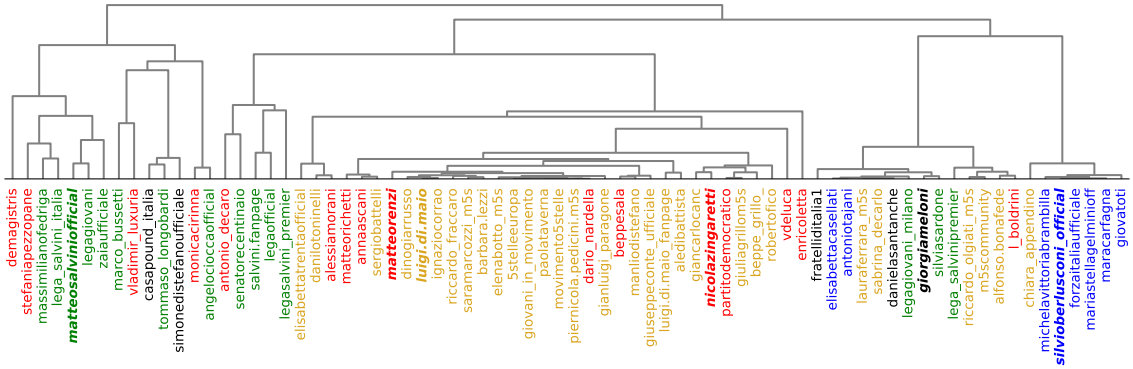
We first focus on the main political leaders of the two countries and study how the communities of co-commenters distribute their interests among their posts. We consider six politicians in each country. Figure 3 shows how the commenters of each community are spread among posts of each politician using a heatmap. Columns represent politicians and rows represent communities. The color of each cell reflects the fraction of the comments of the community members that were published to the posts of the politician.

To gauge similarity of profiles, the top of the heatmaps report the dendrogram that clusters politicians based on their community structure. We define as similarity metric among politicians the Pearson correlation among the activity of communities on their posts. In other words, we compare them by computing the correlation between the corresponding columns of the heatmap, thus used as signatures. As such, two politicians that receive comments from the same communities get a high similarity. The dendrogram clusters together similar politicians based on the communities of their commenters.

Looking at the Brazilian case (Figure 3a), we notice that most communities are interested in a single candidate - Jair Bolsonaro (jairmessiasbolsonaro) - with the large majority of the comments focused on his posts. This behavior is expected given the huge popularity of his posts. Indeed, communities $1 - 9$ comment almost uniquely on Bolsonaro – yet, they result separated. Manual inspection reveals that these commenters comment on different posts. Communities differ mainly for the topics that they comment on as well as for the hour of the day they are active (some commenting on evening posts, others during morning ones). The posts in Figure 4 provide an example of the different topics that communities target. Community 3 comments mostly on posts related



(a) Brazil.



(b) Italy.

Figure 5: Dendogram of political influencers clustered according to commenter communities. Influencers are colored according to their political coalition.

to public events Bolsonaro promoted via Instagram (Figures 4a and Figures 4b), while community 7 comments on posts where the candidate tries to show his proximity with black people in his adolescence to demystify his associations with racism (Figures 4c and Figures 4d).

Focusing on the dendrogram on the top of the figure, Bolsonaro has the highest dissimilarity from the others, i.e., he is the first candidate to be separated from the others. Other clusters reflect accurately the candidates' political orientation. Left-leaning candidates (Ciro Gomes, Fernando Hadaad and Luiz Inacio Lula[5]) are close, as well as the ones leaning towards the right-wing parties (Alvaro Dias, Cabo Daciolo and Jair Bolsonaro).

For the Italian case (Figure 3b), similar considerations hold. Communities $1-10$ focus on Matteo Salvini (matteosalviniofficial). He is the only one for which we identify multiple and well-separated communities. Right-wing party leaders have communities active almost exclusively on their posts, e.g., communities 13 and 14 for Silvio Berlusconi and Giorgia Meloni. Interestingly, other leaders (e.g., Matteo Renzi and Nicola Zingaretti for the Democratic Party and Luigi Di Maio for the Five Star Movement) share a large fraction of commenters in community 11. This suggests these commenters are almost equally interested in the three leaders, who in turn may post content about same topics. Indeed, looking at the dendrogram, these last three profiles are close to each other. Matteo Salvini (leader of the most popular party) has the maximum distance from others. Similar to the Bolsonaro's case, Salvini is a single leader who polarizes communities, thus well-separated from others.

We now broaden our analysis by extending it to all politicians. We aim at studying how communities spread across politicians. To this end, we label each politician according to his/her political *coalition* using available public information.[6] For Brazil, we rely on the Brazilian Superior Electoral Court,[7] while for Italy we use the official website of each party. Rather than reporting the activity of each community on all politicians, we show only the dendrograms that cluster politicians following the same idea as for Figure 3.

Figure 5 shows the results. We use bold to highlight the party leaders/candidates also in Figure 3. The figure clearly shows that politicians of the same parties appear close, meaning that their posts are commented by the same communities. For Brazil, the higher splits of the dendrogram roughly create two clusters, for left-

[5]Fernando Haddad replaced Lula after that Lula was barred by Electoral Justice.

[6]Differently to US or UK, in both Brazil and Italy the political system is fragmented into several parties that form coalitions during and after elections [6].

[7]http://divulgacandcontas.tse.jus.br/divulga/#/estados/2018/2022802018/BR/candidatos

and right-wing parties. In Italy, we can identify three top clusters, reflecting the tri-polar system. Again, this is a natural and expected result. Less expected are the cases in which politicians from distant political leanings attract the interest of the same communities. These cases are put close in the dendrogram. For example, in Italy, we find the profile of Monica Cirinnà (left-wing) very close to Angelo Ciocca (right-wing). Manual inspection reveals a considerable number of disapproving comments to posts of the first politician that are published by commenters supporting the second. The same happens for Vladimir Luxuria - whose some supporters are seen to disapprove Marco Bussetti's posts (and vice-versa). In this case, the structure of the backbone graph and communities reflect the presence of profiles that bridge and blur communities.

In a nutshell, our methodology allows us to highlight the structure of the communities, which reflects people's engagement to political candidates over the spectrum of political orientation. We see that Instagram commenters reflect very well the political orientation of the candidates. Communities are well-shaped around single profiles, and even sub-communities emerge in the case the candidates post content about very different topics. The picture gets fuzzier when commenters of a community act on profiles from the opposite political orientation.

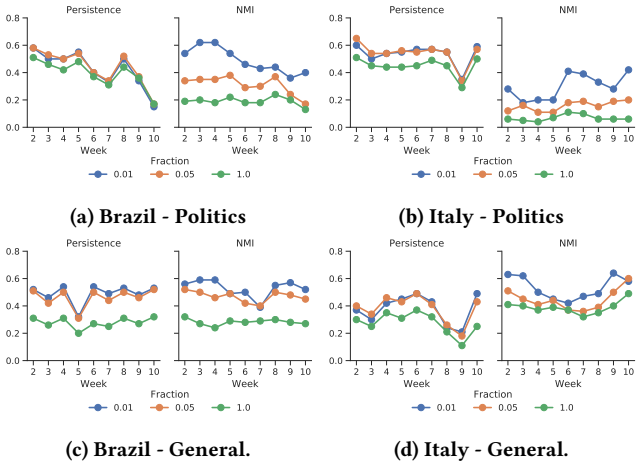
6 RQ2: COMMUNITY EVOLUTION

In this section, we focus on the dynamics of communities during 10 weeks. To this end, we compute two metrics that allow us to measure the evolution of the backbone network and communities over time. First, we compute the *persistence* to measure how the commenters in each backbone graph evolve over time. Given the week w the persistence in w+1 measures the fraction of commenters in the backbone of w who are present in the backbone of $w+1$. If persistence is equal to 1, all commenters in w are still present in $w+1$ (plus eventually others). Yet - their membership to individual communities can change.

To measure the variations in community membership, we compute the *normalized mutual information* (NMI) over the communities [12], taking only commenters who persisted over the two weeks. The NMI ranges from 0 (all commenters changed their communities) to 1 (all commenters remained in the same community). Given two sets of partitions X and Y defining community assignments for nodes in week w and $w+1$, the mutual information of X and Y represents the informational overlap between X and Y, or, in other words, how much we can learn about Y from X (and about X from Y). Let $P(x)$ be the probability that a node picked at random is assigned to community x, and $P(x,y)$ the probability that a node picked at random is assigned to both x in X and y in Y. Let $H(X)$ be the Shannon entropy for X defined as $H(X) = -\sum_x P(x)\log P(x)$. As such, the NMI of X and Y is defined as:

$$NMI(X,Y) = \frac{\sum_x \sum_y P(x,y)\log\frac{P(x,y)}{P(x)P(y)}}{\sqrt{H(X)H(Y)}} \quad (1)$$

To check if the more engaged commenters exhibit a different behavior – i.e., tend to persist more than those who are less engaged – we perform a separate analysis selecting the top-1% and top-5% of commenters among those that entered the highest volume of comments in week w and $w+1$. Next, we compute persistence and NMI restricting to these commenters and compare the results with those obtained with the full set of commenters.



(a) Brazil - Politics (b) Italy - Politics

(c) Brazil - General. (d) Italy - General.

Figure 6: Temporal evolution of commenters in communities. Blue: top 1%, Orange: top 5%, Green: all commenters.

We report results in Figure 6 separately by country and for Politics and General. Considering Politics (Figures 6a and 6b), we note that persistence in Brazil is moderately high, regardless the subsets of commenters. Around 50-60% of commenters remain in the backbone week after week until the first round of elections (week 5). After that we observe a decrease (also due to the drop of commenters in general) until the second round election (week 8), followed by a significant drop after. This trend shows that the commenters were very engaged in the election period, mostly in the first round when the debate included more politicians, senators, congressmen and governors. In the second round fewer candidates faced - yet people were consistently engaged before finally plumbing two weeks after elections. These results corroborate the first intuition we observed in Table 1 – where the number of commenters varied over time.

Considering the membership of commenters within the same community, the NMI shows that the top-1% and top-5% most active commenters (blue and orange curves) are considerably more stable in their communities during the whole time. When checking all commenters in the backbone, we see NMI dropping. This is due to the birth and death of new communities, centered around specific topics, where the debate heats up and cools down. These dynamics attract new commenters that afterward disappear or change the community.

In Italy (Figure 6b), the constant persistence suggests a stable engagement of commenters. We just observe a sudden drop the week after the election, where the interest in the online debate vanished. On the other hand, the NMI is rather low, revealing more dynamicity in community membership - even when we restrict our attention to the most active commenters. Despite commenters in the backbone tend to be the same (persistence > 0.5), they mix among different communities. Considering the result in the perspective of Table 4, we conclude that here we observe a weaker community structure, indicating some degree of overlapping among communities that favor membership changes. This result is also

visible from the dendrogram in Figure 5b, where a lot of influencers receive comments from similar communities.

Moving to the General category, Figures 6c and 6d, we observe similar or lower persistence than in Politics - but it is more stable over time. NMI instead often results in higher for General than Politics, reflecting the better separation between communities, which persist over time. More in detail, for Brazil (Figure 6c) we observe that persistence and NMI are high and stable – especially for the most active users. This result suggests that the most engaged commenters have diverse, very specific, and stable interests. Again, it reflects the high heterogeneity of posts and influencers in the General category. Moving to Italy, Figure 6d shows that persistence is small and varies over time. Here the lower popularity of Instagram in Italy than in Brazil may play a role, coupled with the smaller number of comments (compare with Table 1). However, NMI is high and stable. We conclude that despite many users that do not persist, the remaining ones are very loyal to their interests.

In a nutshell, people commenting in politics are more volatile, with debates suddenly growing and cooling down, e.g., after elections. On the contrary, in General topics, we observe more consistent and well-separated communities, caused by the different interests in following influencers from different sectors.

7 CONCLUSIONS

We analyzed the community structure in the network of Instagram commenters. To this end, we designed a methodology to extract salient interactions between commenters based on a null model that allows us to capture unexpected interactions. We then generated a backbone network that better captures interactions with large evidence of coordination. We then applied community detection algorithms to obtain relevant groups of commenters.

We applied this methodology to a dataset containing the Instagram activity on several public profiles during the electoral periods of Italy and Brazil, comparing it to profiles of general influencers. The communities obtained for political and general profiles present strong structural differences and distinct dynamics during the electoral periods. We observed, for example, weaker communities in the political scenario, which however create more intense debate around elections.

Future work includes a deeper look into highly coordinated behavior on Instagram, coupled with text analysis of comments by each community, e.g., to understand whether and how communities of coordinated commenters promote or demote particular profiles or focus on certain topics.

ACKNOWLEDGMENTS

The research leading to these results has been funded by the SmartData@PoliTO Center for Big Data technologies at Politecnico di Torino, Brazilian National Council for Scientific (CNPQ), Technological Development and Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) and Minas Gerais State Foundation for Research Support (FAPEMIG).

REFERENCES

[1] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences* 115 (2018), 11221–11230.
[2] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2018. Sequences of sets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008.
[4] M. Coscia and F. M. H. Neffke. 2017. Network Backboning with Noisy Data. In *IEEE 33rd International Conference on Data Engineering (ICDE)*. 425–436.
[5] Carlos Henrique Gomes Ferreira, Breno de Sousa Matos, and Jussara M Almeira. 2018. Analyzing dynamic ideological communities in congressional voting networks. In *International Conference on Social Informatics*. Springer, 257–273.
[6] Carlos H G Ferreira, Fabricio Murai, Breno de Souza Matos, and Jussara M de Almeida. 2019. Modeling Dynamic Ideological Behavior in Political Networks. *The Journal of Web Science* 7 (2019).
[7] Katerina Gorkovenko and Nick Taylor. 2017. Understanding How People Use Twitter during Election Debates. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*.
[8] Derek Greene, Dónal Doyle, and Padraig Cunningham. 2010. Tracking the Evolution of Communities in Dynamic Social Networks. *Proceedings of the International Conference on Advances in Social Network Analysis and Mining, ASONAM 2010* 2010.
[9] Yili Hong. 2013. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis* 59 (2013), 41–51.
[10] Roope Jaakonmäki, Oliver Müller, and Jan Vom Brocke. 2017. The impact of content, context, and creator on user engagement in social media marketing. In *Proceedings of the 50th Hawaii international conference on system sciences*.
[11] Seungbae Kim, Jinyoung Han, Seunghyun Yoo, and Mario Gerla. 2017. How are social influencers connected in instagram?. In *International Conference on Social Informatics*. Springer.
[12] Andrea Lancichinetti, Santo Fortunato, and János Kertész. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* 11, 3 (2009), 033015.
[13] Matteo Manca, Ludovico Boratto, Victor Morell Roman, Oriol Martori i Gallissà, and Andreas Kaltenbrunner. 2017. Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. *Online Social Networks and Media* 1 (2017), 56–69.
[14] Caroline Lego Munoz and Terri L Towner. 2017. The image is the message: Instagram marketing and the 2016 presidential primary season. *Journal of Political Marketing* 16, 3-4 (2017), 290–318.
[15] MEJ Newman. 2018. Network structure from rich but noisy data. *Nature Physics* 14 (2018), 542–546.
[16] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69 (2004), 026113.
[17] Josep M Pujol, Vijay Erramilli, and Pablo Rodriguez. 2009. *Divide and Conquer: Partitioning Online Social Networks*. Technical Report arXiv:0905.4918.
[18] Gustavo Resende, Philipe Melo, Julio CS Reis, Marisa Vasconcelos, Jussara M Almeida, and Fabrício Benevenuto. 2019. Analyzing textual (mis) information shared in WhatsApp groups. In *Proceedings of the 10th ACM Conference on Web Science*.
[19] Lei Shi, Hanghang Tong, Jie Tang, and Chuang Lin. 2015. Vegas: Visual influence graph summarization on citation networks. *IEEE Transactions on Knowledge and Data Engineering* 27 (2015), 3417–3431.
[20] Tanase Tasente. 2020. The #Brexit on the Facebook pages of the European institutions. *Technium Social Sciences Journal* 3, 1 (2020), 63–75.
[21] Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, Fabricio Murai, Flavio Figueiredo, Ana Paula Couto da Silva, and Jussara M Almeida. 2019. Towards Understanding Political Interactions on Instagram. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*.
[22] Yuan H Wang. 1993. On the number of successes in independent trials. *Statistica Sinica* (1993), 295–312.
[23] Zhenhua Wu, Lidia A Braunstein, Shlomo Havlin, and H Eugene Stanley. 2006. Transport in weighted networks: partition into superhighways and roads. *Physical review letters* 96 (2006).
[24] K. Zarei, R. Farahbakhsh, and N. Crespi. 2019. Typification of Impersonated Accounts on Instagram. In *IEEE 38th International Performance Computing and Communications Conference (IPCCC)*.
[25] Wei Zeng, Meiling Fang, Junming Shao, and Mingsheng Shang. 2016. Uncovering the essential links in online commercial networks. *Scientific reports* 6 (2016), 34292.
[26] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020).