Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120515

Identifying Influential Users on Twitter: A Case Study from Paris Attacks

Layal Abu Daher[1,*], *Islam Elkabani*[1] *and Rached Zantout*[2]

[1] Department of Mathematics and Computer Science, Beirut Arab University, Debbieh, Lebanon
[2] College of Engineering, Rafik Hariri University, Mechref, Lebanon

Received: 28 Jun. 2018, Revised: 2 Jul. 2018, Accepted: 6 Aug. 2018
Published online: 1 Sep. 2018

Abstract: Due to the spread of technology and world wide web, online social media has invaded every home in the world; hence, the analysis of such networks is considered an important yet challenging case of study for researchers. One of the most interesting fields of study in social network analysis is identifying influential users who are important actors in online social networks by having an impact on others. This work investigates the problem of identifying influential users on Twitter. Since Twitter is a user-friendly interactive platform, it is now an apparent competitor to other social medias as far as user interaction. Twitter is browsed by a variety of users, the most important are the most influential ones among them all. In order to identify influential users, a data set is collected between December 2015 and March 2016 reflecting real tweets from the top trendy hashtags on Twitter. In this paper, different measures are used such as influence measures, centrality measures and activity measures. In addition, association learning has been used to detect relationships between users. After identifying the influential users from association learning, these influential users are compared to the results of the abovementioned measures. The results of this study indicate that identifying influential users from association learning and validating these identified users with the results of influence measures is an effective method for detecting the influence of users on online social networks.

Keywords: Association rule learning, hashtags, influential users

1 Introduction

One of the fastest ways for sharing and propagating information nowadays is the Online Social Networks (OSNs). Users on such platforms share opinions, feelings, moods, life experiences, and even needs. Topologically, the social network graph consists of distributed nodes connected by edges to form relationships. These relationships are either directed or undirected depending on the social network graph. For instance, the famous Twitter [1] is considered to be a directed graph since the relationship between two users does not need acceptance from both sides. However, a user might follow another user who in turn, does not have to follow the first user back knowing that on Twitter, only 22% of relationships are reciprocal [2]. Users on Twitter initiate hashtags while writing a certain post. Hashtags are a group of characters preceded by # symbol. According to hashtags, tweets are categorized based on subjects like political, artistic, athletic or even entertaining. For example, *#ParisAttacks*, was one of the trendiest hashtags in the period between December 2015 and March 2016. Users created many hashtags where they tweeted and retweeted about different subjects related to the above hashtag. Some of these initiated hashtags remained in use in users postings over the whole period of time, while others faded by time. It is interesting to study why different hashtags have different lifetimes and whether users influence others through using a hashtag or is it the subject of the hashtag which influences users to tweet using the hashtag. In OSNs, the users are the most important factors, and identifying or predicting the influential users among billions of them can be very helpful in predicting the future of these networks. In this paper, a study is done on a dataset of some trendy hashtags to identify influential users associated with the hashtags. The study has been conducted from two different perspectives: Social Network Analysis (SNA) perspective and Data Mining (DM) perspective, specifically Association Rule Learning (ARL). For evaluation purposes, comparison between the results of both perspectives is done. In the SNA

[*] Corresponding author e-mail: layal.abudaher@bau.edu.lb

perspective, we adopt, from the literature, two Topological Measures (TM), specifically, the Betweenness Centrality (BC) and Closeness Centrality (CC) in addition to three Activity Measures (AM), namely, General Activity (GA), Topical Signal (TS) and Signal Strength (SS). In addition, we introduce a new topological measure, namely Jaccard Weighted In-degrees(JWI) as a better topological measure indicator. The three TM along with the three AM were used to study the indirect influence of users' topology and activity level on identifying influential users in the hashtags under study. In order to set a standard to compare with for identifying the users' influence based on the previously-described measures, we introduce two novel direct influence measures, namely Count of Posting Followers (CPF) and Users Being Retweeted (UBR). From the data mining perspective, ARL data mining has been adopted to identify influential users and predict user participation in the hashtags under study. For evaluation purposes, a comparative study is then conducted between the results from both perspectives where conclusions are drawn.

2 Related Work

Browsing social media is one of the routines done by users more than once on a daily basis since these networks have evolved to become the main source of information for a large part of our society. The interest of many people in OSNs pushed researchers to discover and investigate social networks from all perspectives such as evolution, dynamicity, prediction of users participation, detection of influential nodes and measuring user influence.

A study done in [2] examined the first comprehensive study about Twitter social network. After 7 years, another study [3] re-examined Twitter social network. The results of the re-examination revealed that Twitter social network has gained popularity and its usability increased more by 10-fold. Moreover, reciprocal relationships between users also increased, however, 12.5% of users in the year 2009 have left Twitter. It was also noticed that the network connectivity between users has decreased. Furthermore, popularity of some users also changed where non-popular users became popular and influential. The re-examination of Twitter social network proved that the dynamics of Twitter is very high and the relationships between different users is constantly changing leading to a change to the whole network.

In [4], the authors studied the stability of groups by determining their dynamicity over time focusing on the diversity of members and their social activities. The dataset was downloaded from both World of Warcraft (WOW) which is an online game and DBLP which is a database-containing information on many journals and conferences related to the computer science field. The authors extracted a set of features that describe group composition, activities within the group and structural aspects of a group for each dataset. They used six different classification methods (ZeroR baseline, Nave Bayes, Decision Stump, J48 decision tree, Bagging and RandomForest) the highest percentage of accuracy, 84.78%, was obtained when (RandomForest) was applied to the WOW dataset. For the DBLP dataset, the percentage of accuracy reached 90.55% achieved by (Bagging) classifier. The results of this study showed that it is possible to predict group stability with high accuracy using a range of features.

In [5], experimental studies were done on four social networks for predicting group evolution. The GED method (Group Evolution Discovery) was used from [6] to discover group evolution in the social network. The inclusion measure is the most important measure in GED method which allows the inclusion of one group in another. GED method provides a balance between both the quantity and quality of group members. The key members were determined using different centrality measures such as centrality degree, betweenness degree, page rank and social position. The dataset was collected from four social networks: Wroclaw University of Technology email communication, salon24, which is a portal that contains political discussions, the Enron e-mail network and the portal extradom.pl, which gathers people interested in building their own houses in Poland. Experiments were performed using a data mining software called WEKA where ten different classifiers were utilized for each social network. These classifiers are: BayesNet, NaiveBayes, IBk, KStar, AdaBoost, DecisionTable, JRip, ZeroR, J48 and RandomForest. The results of this study show that the classifier that uses an input of several preceding groups sizes and events, produces very promising results. The experimental analyses on the four datasets collected for this study revealed that the two classifiers, DecisionTrees and RandomForest, provided the most accurate results. Additionally, the GED method used for change identification can be successfully used as a good indicator.

Identifying the most influential users in an online social network has been recently a very interesting field of study in Social Network Analysis (SNA). For example, in [7], the focus was on comment mining, in [8] concentration was on predicting information cascade. In addition, there has been a very deep concentration on studying centrality measures as an important indicator for measuring the degree of influence of a user in a social network [9]. Centrality measures assist in identifying the important users in a social network from a topological perspective. Some of the centrality measures used in the literature are: Closeness Centrality [10], Betweenness Centrality [11], Eigenvector Centrality [12] and more. Other types of measures concentrate on the degree of activity of users as an indicator of their influence on such social networks. The activity of the users can affect their connectedness, influence and popularity [13].

In [14], a large amount of data was gathered from Twitter to analyze users' influence. The authors compared three different measures of influence: indegree, retweets, and mentions. They also focused on different topics, where they examined how the three different influence measures performed in spreading popular news topics. They also inspected the dynamics of an individual's influence by topic and over time. Furthermore, they characterized specific behaviors that make ordinary individuals gain high influence over a short period of time. The findings revealed from this study have direct effects not only on the design of social media, but also on viral marketing. The choice of analyzing the abovementioned three diverse influence measures provided better understanding of the different roles users play in social media. Indegree represents popularity of a user; retweets represents the content value of one's tweets; and mentions represents the name value of a user. A variety of influence across topics is revealed, where this variety could be very effective in the advertisement in Twitter if one is to employ influential users. The analysis showed that most influential users hold significant influence over a variety of topics. Moreover, ordinary users were found to gain influence by focusing on a single topic and posting creative and insightful tweets that are perceived as valuable by others, as opposed to simply communicating with others.

In [15] the authors worked on detecting influential users based on performed actions (e.g. comments or likes) on posts in Facebook pages. ARL was used in order to predict how users are getting connected with each other based on posts, comments and likes on groups on Facebook. The prediction is done based on the activeness of users within posts with similar topics. 2,443 active users interacting on 610 posts with a total of 14,117 comments were extracted. Almost 5,000 association rules were generated with high confidence of correctness 95%. These rules were proven to be dependent on the active users, via the lift metric. This proves that identifying influential users using ARL can be done.

3 Approach

In this study, the problem of identifying influential users is addressed from two different perspectives: SNA perspective and DM perspective. In the SNA perspective, and as illustrated in figure 1, the hashtags for analysis are selected then data of the users posting on the selected hashtags is collected. Following that, ordered (descending) lists of users are created based on the two direct influence measures and the six indirect influence measures. The eight-ordered lists are then used to identify the top ranked users from the six indirect influence measures and compared to the ones identified from the two direct influence measures.

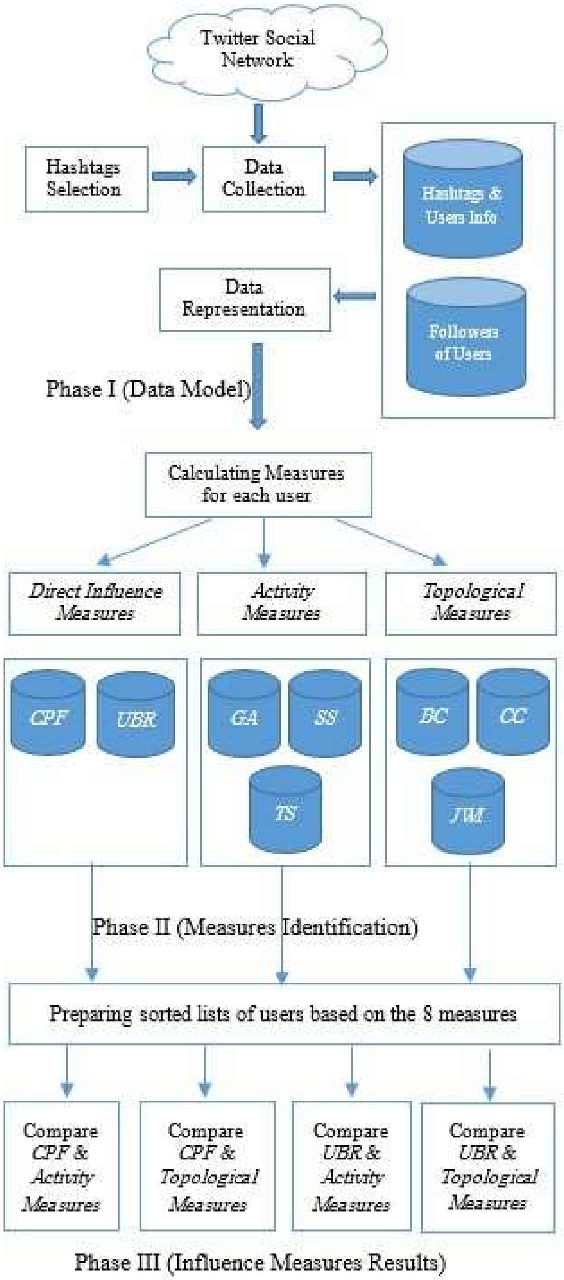
Twitter Social Network
Hashtags Selection
Data Collection
Hashtags & Users Info
Followers of Users
Data Representation
Phase I (Data Model)
Calculating Measures for each user
Direct Influence Measures
Activity Measures
Topological Measures
CPF
UBR
GA
SS
TS
BC
CC
JW
Phase II (Measures Identification)
Preparing sorted lists of users based on the 8 measures
Compare CPF & Activity Measures
Compare CPF & Topological Measures
Compare UBR & Activity Measures
Compare UBR & Topological Measures
Phase III (Influence Measures Results)

Fig. 1: Diagram of SNA Perspective

In the DM perspective, and as depicted in figure 2, ARL is conducted by first preprocessing of data so that input files for the FP-Growth algorithm contained on the left hand side tweeters who influence the retweeters on the right hand side. After the preprocessing phase, the FP-Growth algorithm is applied and rules are built. These rules are post-processed to remove duplicate rules. Finally, a set of influential users are identified from the built rules for the three hashtags under study. A comparative study is then conducted to see whether the influential users identified using the association rules match the top ranked users in the SNA perspective.

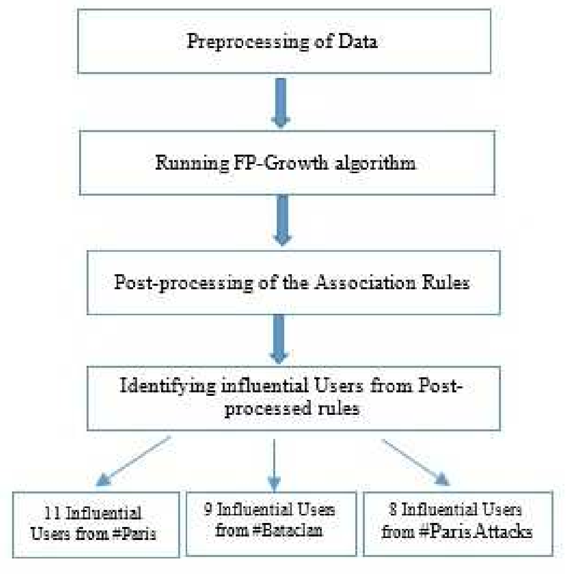
Preprocessing of Data
Running FP-Growth algorithm
Post-processing of the Association Rules
Identifying influential Users from Post-processed rules
11 Influential Users from #Paris
9 Influential Users from #Bataclan
8 Influential Users from #Paris Attacks

Fig. 2: Diagram of DM Perspective

4 Data Model

The dataset used in this analytical study constitutes of hashtags selected from the list of the trendiest hashtags in Twitter according to hashtagify.me website [16]. The information was collected during the period between December 2015 and March 2016. This website provides a list of the trendiest hashtags. In addition, it also offers the hashtags correlated to each hashtag in the trendiest hashtags list. The correlated hashtags concurrently occur with the associated hashtag in the same post. In the list of the trendiest hashtags in the abovementioned duration, #ParisAttacks existed. The ten correlated hashtags for this trendy hashtag as well as their percentage of correlation are listed in Table 1.

The activity on the hashtag #ParisAttacks and its correlated hashtags is recorded during the aforementioned period. Recording has started after the number of posting users (tweeters and retweeters) stabilized.

4.1 Data Collection

The data in the selected hashtags are gathered using the crawler designed in [17]. This crawler collects tweets and retweets on a certain hashtag and their metadata. The information collected includes the text of the tweet, the text of the retweet and the count of followers of the tweeters or retweeters. Another crawler is implemented to collect the list of followers of the posting users. The data collected from the trendiest hashtag and its correlated hashtags consist of 24,026 tweets and 27,928 retweets posted by 28,613 unique users as shown in Table 2. 4,919 out of 28,613 (17.2%) of users tweeting or retweeting on these hashtags has private profiles. This is why they are removed from the dataset in the preprocessing phase. The corpus selected for the analytical study is based on three hashtags which are #ParisAttacks, #Paris, and #Bataclan. The first hashtag, #ParisAttacks, is chosen because it is the root hashtag and it is directly related to the terrorist attacks that occurred in Paris in November 2015. The second hashtag, #Paris, is chosen for analysis because, according to Table 1, it has the highest percentage of correlation with the trendiest hashtag #ParisAttacks and the highest number of posting users as shown in Table 2. Although #PrayForParis and #rechercheParis have highest percentage of correlation with #ParisAttacks after #Paris as shown in Table 1, the total number of posting users in these two hashtags is less than that of #Bataclan, as illustrated in Table 2. This is why we chose #Bataclan to be the third hashtag for analysis, taking into account that Bataclan is the name of the theatre where the terrorist attack occurred.

Table 1: The %of the 10 correlated #s to #ParisAttacks

#ParisAttacks	
Hashtags	Correlation
#Paris	7.5%
#PrayForParis	5.2%
#rechercheParis	3.5%
#Bataclan	3.1%
#ISIS	2.9%
#Prayers4Paris	1.9%
#France	1.7%
#fusillade	1.7%
#ParisShooting	1.3%
#Iran	1.3%

4.2 Data Representation

The dataset collected is modeled as a directed graph. The nodes represent distinct posting users and the links represent the relationships between the users. The link between two nodes constitutes of a source node, a target node, and a weight of the link between the source and target nodes. A link is established between user A and user B if user A is found in the list of followers of user B. The weight of the link between nodes A and B is the

Table 2: Tweets and Retweets in #ParisAttacks

#ParisAttacks		
Hashtags	Tweets	Retweets
#ParisAttacks	1210	2330
#Paris	5679	6973
#PrayForParis	1504	711
#rechercheParis	1	19
#Bataclan	1024	2513
#ISIS	3324	6730
#Prayers4Paris	5	50
#France	3849	2322
#fusillade	79	530
#ParisShooting	25	16
#Iran	7326	5734

Jaccard's similarity coefficient [18] displayed in Equation 1.

$$J(A,B) = \frac{|followers(A) \cap followers(B)|}{|followers(A) \cup followers(B)|} \tag{1}$$

where:

- –followers(x) is the list of followers of user x
- –The numerator of Equation (1) is the number of followers of both users A and B
- –The denominator of Equation (1) is the number of followers in the union of the followers of both users A and B

Gephi [19], a software for graph analysis, was used to conduct the calculations of the centrality measures.

5 Influence Measures

Two direct influence measures are used in this study to complement the classical activity measures and the topological measures (indirect influence measures). Moreover, a topological measure is also used in addition to the centrality measures available in the state of art in order to determine a user's level of social connectedness to users posting on the same hashtags.

5.1 Direct Influence Measures

Most of the applications and tools in the state of art compute the user influence from the count of followers/friends, tweets, retweets, mentions and replies. In this paper, the two new direct influence measures introduced in [20] are also used. Namely, the CPF and UBR. The two new measures are expected to help study the direct influence of users on the dynamicity of hashtags on Twitter. CPF represents the count of total followers of user A who are posting on the same hashtag as represented in Equation 2.

$$CPF = \sum_{i \in F(A)} x_i \tag{2}$$

where:

- –F(A) is the set of followers of User A
- –x$_i$ = 0 if user i, a follower of user A, does not post
- –x$_i$ = 1 if user i, a follower of user A, posted

UBR represents the sum of the retweets done on the tweets of user A as represented in Equation 3.

$$UBR = \sum_{t \in T(A)} RT(t) \tag{3}$$

where:

- –T(A) is the set of tweets by user A
- –RT(t) is the number of retweets of a tweet t

The idea behind the CPF is that a user whose followers are posting more on his same hashtag is expected to have had a direct influence on those users to post using that particular hashtag. The idea behind UBR is that a user whose posts are retweeted the most is expected to have had a direct influence on increasing the number of users posting using hashtags which the user had used in his original tweets.

5.2 Activity Measures

On Twitter, a user is active when he/she participates in real tweets and retweets. The General Activity (GA) measure of a user [21] is calculated by summing the number of tweets and retweets of a certain user A on a certain hashtag on Twitter and as illustrated in Equation 4.

$$GeneralActivity(A) = T + RT \tag{4}$$

where:

- –T represents the count of all tweets of user A
- –RT represents the count of all retweets of user A

Topical Signal (TS) [22] is the second activity measure used. TS is the ratio of the general activity of user A and the total number of tweets posted on a certain hashtag as shown in Equation 5.

$$TS(A) = \frac{(T+RT)_{|specifichashtags}}{TotalNumberofTweets} \tag{5}$$

The third activity measure used is Signal Strength (SS) [22], which identifies the strength of the TS. SS is the division of the tweets of user A by the total number of his tweets and retweets on a certain hashtag on Twitter as presented in Equation 6.

$$SS(A) = \frac{T}{T+RT} \tag{6}$$

5.3 Topological Measures

Two classical topological measures, namely the Betweenness Centrality (BC) and Closeness Centrality (CC) measures are used in this study. These two measures are used to study how the centrality of users on certain hashtags can affect the evolution of these hashtags.

5.3.1 Betweenness and Closeness Centrality Measures

BC is the first centrality measure adopted in this study. BC is a centrality measure in graph theory based on the shortest paths. In a connected weighted graph, for every pair of nodes, there exists at least one shortest path between the nodes such that the sum of the weights of the edges is minimized. The BC for each node is the number of these shortest paths that pass through the node. The reason for choosing this centrality measure is that the node with higher BC would have more control over the network, because more information passes through it.

The second centrality measure adopted from the literature is the Closeness Centrality (CC). CC is a degree measure to which a node is near all other nodes in a network. This centrality measure is based on the length of the shortest paths from a node to all other nodes in the network. It measures the visibility or accessibility of each node with respect to the entire network.

5.3.2 Jaccard Weighted In-degrees

Unlike centrality measures, Jaccard Weighted In-degrees (JWI) measures the social connectedness of the users posting on a social network graph more than measuring the effect of the user in a network graph. This means that, JWI measures the density of the relationships between users more than measuring the effect of a certain user on other users in a network. As displayed in Equation 7, and in order to calculate JWI measure, we add the Jaccard's similarity weights of all the edges directed to user A.

$$JWI(A) = \sum_{i=1}^{n} w_i(A) \tag{7}$$

where wi is the weight of link i ending on node A.

6 Association Rule Learning

The method of matching items in different transactions is called ARL. This method aims to find out how items affect each other by analyzing how frequently items appear together in a specific dataset [23]. This is done by considering two criteria, namely, support and confidence. The support of the association rule indicates the frequency of items in a specific dataset, while the confidence of the association rule indicates how reliable the rule is. The higher the value of confidence, the more likely the items occur in a specific dataset [24].

6.1 Evaluation Metrics

Several evaluation metrics exist for association rule learning. The first metric, support, is an indication of how frequently a set of users appear in the list of posts D in a certain hashtag. As represented in Equation 8, support measure is the ratio of the occurrences of a set of users on a certain hashtag by the total number of posts in a certain hashtag.

$$support(\{A,B\}) = \frac{\{A,B\}}{|D|} \tag{8}$$

The confidence displayed in Equation 9 designates the set of posts where users A and B are participating and where user C also participates. For example, three users A, B and C are concurrently posting on four common posts, whereas user A and B are concurrently posting eight posts. This yields a confidence of 4/8 = 0.5. This means that there is a 50% confidence that user C participates in the same posts where users A and B are posting.

$$confidence(\{A,B\} \rightarrow C) = \frac{support(\{A,B,C\})}{support(\{A,B\})} \tag{9}$$

Lift is the third evaluation metric of association rules used in this study. This measure displays the ratio of interdependence of values displayed in equation 10. When the value of lift reaches 1, this means that the users in the existing rules are independent. As the value of lift exceeds 1, this means that the users are dependent of each others.

$$lift(\{A,B\} \rightarrow C) = \frac{support(\{A,B,C\})}{support(\{A,B\}) \times support(\{C\})} \tag{10}$$

The abovementioned measures constitute the metrics of ARL. As noticed from the equations of these metrics, the higher the values, the higher the relevance for prediction.

6.2 Building the Association Rules

In order to build the association rules from the dataset, several association rule learning algorithms were evaluated. The Apriori Algorithm (AA) [25] proved to be an effective algorithm for ARL. However, running this algorithm on the available dataset showed serious scalability and memory exhaustive issues. The FP-Growth, an improvement on the AA, was designed to eliminate some of AA disadvantages such as memory usage and runtime [26]. FP-Growth works by first counting all the users who tweeted in a certain hashtag. After setting the appropriate threshold for the three metrics, support, confidence and lift, a sorted list of users who tweeted is created. The list is sorted according to the count of occurrence of each user. Then, a tree based on the sorted list is created and association rules are created for every branch of the tree based on the predefined threshold values. In order to identify the threshold values for the dataset under consideration, different threshold values of the three metrics mentioned above are tested until association rules are successfully created. The representation of a built rule itself is divided into two parts, left hand side and right hand side. For example the rule $\{u1, u2\} \rightarrow$ {u3, u4} indicates that the left-hand-side users u1 and u2 influence the right-hand-side users u3 and u4, i.e., when the left-hand-side users are both active on a post, the right-hand-side users will consequently be active.

7 Influence Measures: Experiments & Results

After calculating the direct and indirect influence measures described in section 5 for the three hashtags under study, ordered (descending) user lists are created for each measure. The comparison between the different ranked measures is conducted based on the following:

- Comparing top ranked users from both CPF and AM.
- Comparing top ranked users from both CPF and TM.
- Comparing top ranked users from both UBR and AM.
- Comparing top ranked users from both UBR and TM.

For the three selected hashtags under study, the intersection results of CPF with the level of the three AM, namely GA, TS and SS are presented in Table 3. The results do not exceed 34% in #ParisAttacks and 27% in #Paris and #Bataclan. This can be explained by the observation that the activity of a certain user A posting on a hashtag does not necessarily influence his followers to post on the same hashtag but rather for other reasons like being interested in an event.

Table 3: CPF and AM similarity

CPF Intersection Results						
	#Paris		#ParisAttacks		#Bataclan	
Activity Measures	25% ∩ (1943 Users)		25% ∩ (497 Users)		25% ∩ (532 Users)	
	No.	%	No.	%	No.	%
GA	525	27	170	34	141	27
TS	525	27	170	34	141	27
SS	298	15	142	29	142	27

Table 4 presents the results of intersection between CPF and the three TM, namely BC, CC and JWI. The results of intersection are between 48% and 65% in #Paris and #ParisAttacks, while it do not exceed 27% in #Bataclan. This indicates that the TM of a user may not be a good measure of the users influence on his/her followers as it depends on the actual hashtag.

Table 4: CPF and TM similarity

CPF Intersection Results						
	#Paris		#ParisAttacks		#Bataclan	
Topological Measures	25% ∩ (1943 Users)		25% ∩ (497 Users)		25% ∩ (532 Users)	
	No.	%	No.	%	No.	%
Betweenness	1052	54	298	60	118	22
Closeness	927	48	263	53	122	23
JWI	1255	65	320	64	145	27

Table 5 presents the intersection results of UBR with the level of the three AM. Results of intersection with GA and TS are between 74% and 96% in #Paris and #ParisAttacks, while SS intersection with UBR reach 52%. Whereas, in #Bataclan, the intersection results of UBR with GA and TS are 58%, while in SS results reach 97%. This indicates that in #Paris and #ParisAttacks, the GA of a user has more influence on the user being retweeted than the influence of the user' original tweets. However, in #Bataclan, the posted tweets of a user is more influential. This indicates that users in #Bataclan are being retweeted more due to the content of their posts on #Bataclan. However, in #Paris and #ParisAttacks, as the users add more posts on these hashtags, they are retweeted more.

Table 5: UBR and AM similarity

UBR Intersection Results						
	#Paris		#ParisAttacks		#Bataclan	
Activity Measures	25% ∩ (1943 Users)		25% ∩ (497 Users)		25% ∩ (532 Users)	
	No.	%	No.	%	No.	%
GA	1439	74	408	82	310	58
TS	1439	74	478	96	310	58
SS	778	40	256	52	514	97

In table 6, the results of intersection between UBR and the three TM are displayed. The results do not exceed 43% in #Paris and #ParisAttacks while the JWI reached 68% in #Bataclan. This indicates that the users who are retweeted more, are the more socially connected users in #Bataclan, but this is not the case in #Paris and #ParisAttacks. This indicates that in familiar hashtags, users do not have to be connected in order to be retweeted, however in non-familiar hashtags like #Bataclan, users have to be connected in order to know that their connections are posting on such hashtags. Since most of the intersection results linearly increase with the percentage of top users, outliers exist. The reason for this is that many users have had the same rank in the top 50% and 75%. This is why our focus was on the top 25% in this analytical comparison.

Table 6: UBR and TM Similarity

UBR Intersection Results						
	#Paris		#ParisAttacks		#Bataclan	
Topological Measures	25% ∩ (1943 Users)		25% ∩ (497 Users)		25% ∩ (532 Users)	
	No.	%	No.	%	No.	%
Betweenness	586	30	169	34	167	31
Closeness	548	28	165	33	149	28
JWI	845	43	172	35	361	68

8 Association Rules: Experiments & Results

Using the FP-Growth algorithm described in section 6.2, and after post-processing, different numbers of association rules are obtained depending on the hashtag. In #Bataclan, there are 376 distinct posting users and 12 rules are created. In #ParisAttacks, there are 661 distinct posting users and 46 rules are created. In #Paris there are 1134 distinct posting users and 4107 rules are created. It is noticed that the number of created rules for #Paris is the highest. This is because #Paris has the highest number of distinct posting users. In addition, those users post generally about everything related to Paris and their posts are not restricted to the terrorist attacks which occurred in Bataclan Theater.

Table 7 shows descriptive statistics of all the remaining rules after the preprocessing phase of the three hashtags under study. In the three hashtags under study, the mean of the confidence of the rules created is relatively high and the standard deviation is low. Moreover, the mean lift values in the three hashtags indicate a strong dependency of the users within a rule due to lift high values. From all the learned rules in each hashtag under study, it is noticed that 8 influential users are identified for #ParisAttacks, 11 for #Paris and 9 for #Bataclan.

Table 7: Statistics of computed association rules

#ParisAttacks			
Evaluation Metric	Mean	Median	Std.
Support	6.85	4	8.53
Confidence	0.76	0.93	0.27
Lift	122.68	101.9	96.70
#Paris			
Evaluation Metric	Mean	Median	Std.
Support	6.99	7	0.051
Confidence	0.99	1	0.02
Lift	208.01	208.29	5.68
#Bataclan			
Evaluation Metric	Mean	Median	Std.
Support	3.08	3	0.29
Confidence	0.71	0.75	0.22
Lift	80.76	82.83	36.07

9 Discussion and Findings

The goal of this analytical study is to see how influential users identified from the ARL match the top users identified from the influence measures described in section 5. Therefore, sorted lists for each measure from the direct influence measures (CPF and UBR) and the indirect influence measures (GA, TS, SS, BC, CC and JWI) are prepared. Subsequently, the influential users detected from the ARL are ranked based on the ordered lists prepared. A user in the top 1% of the measure receives a measure score of 1. This means that the user is more influential than 99% of the users according to this measure. Tables 8 to 10 present the results of ranking users in the three hashtags under study.

9.1 Discussion

The statistical results presented in Table 8 reveal that, in #ParisAttacks, the eight influential users identified from ARL when ranked in the ordered list of CPF, have a ranking mean of 45.6%, a median of 33% and a standard deviation of 35.5%. This indicates that these influential users do not influence their followers to post on #ParisAttacks. In ranking the influential users in the ordered list of UBR, the ranking mean is 2.5%, the median is 1.5% and the standard deviation record 1.9%. This clearly indicates that the influential users identified from ARL are among the top ranked users being retweeted. Moreover, the results of GA record values of 2.1%, 1% and 1.8%, TS and SS results record 6.8%, 1% and 10.1% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are obviously active on #ParisAttacks. Regarding the results of the TM, the results of BC record values of 43.1%, 30% and 38.5%, CC results record 45.8%, 43.5% and 35.8% and JWI results record 40.2%, 33% and 26.4% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are not among the central users in #ParisAttacks, however, their social connectedness represented by JWI indicat better results than other centrality measures.

Concerning the statistical results of #Paris presented in Table 9, the eleven influential users identified from ARL when ranked in the ordered list of CPF show that the

Table 8: Descriptive statistics of rating influential users from ARL with influence measures in #ParisAttacks

#ParisAttacks			
Influential users from ARL with Influence Measures	Mean	Median	Std.
Users from ARL with CPF	45.6	33	35.5
Users from ARL with UBR	2.5	1.5	1.9
Users from ARL with GA	2.1	1	1.8
Users from ARL with TS	6.8	1	10.1
Users from ARL with SS	6.8	1	10.1
Users from ARL with BC	43.1	29.5	38.5
Users from ARL with CC	45.8	43.5	35.8
Users from ARL with JWI	40.2	33	26.4

ranking mean record a value of 43.7%, the median 36% and the standard deviation is 29.3%. This indicates that these influential users do not influence their followers to post on #Paris. The ranking mean of the influential users in the ordered list of UBR is 19.4%, the median is 1% and the standard deviation record 39.9%. This indicates that the influential users identified from ARL are among the top ranked users being retweeted especially due to the value of the median which record 1. Concerning the statistical results of the activity measures, the results of GA and TS record values of 19.2%, 1% and 39.9%, whereas SS results record 34.5%, 29% and 30.1% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are obviously active on #Paris. However, the results of the SS measure indicate that these influential users are retweeting other than posting original tweets. Regarding the results of the TM, BC record values of 43.7%, 43% and 32.6%, CC record 42.6%, 48% and 34.8% and JWI results record 69.5%, 71% and 22.8% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are not among the central and the socially connected users in #Paris.

Table 9: Descriptive statistics of rating influential users from ARL with influence measures in #Paris

#Paris			
Influential users from ARL with Influence Measures	Mean	Median	Std.
Users from ARL with CPF	43.7	36	29.3
Users from ARL with UBR	19.4	1	39.9
Users from ARL with GA	19.2	1	39.9
Users from ARL with TS	19.2	1	39.9
Users from ARL with SS	34.5	29	30.1
Users from ARL with BC	43.7	43	32.6
Users from ARL with CC	42.6	48	34.8
Users from ARL with JWI	69.5	71	22.8

The statistical results presented in Table 10 show that, in Bataclan, the nine influential users identified from ARL when ranked in the ordered list of CPF, record a ranking mean of 49.8%, a median of 46% and a standard deviation of 34.2%. This indicates that these influential users do not influence their followers to post on #Bataclan. Furthermore, ranking of the influential users in the ordered list of UBR, the ranking mean is 3.3%, the median is 1% and the standard deviation record a value of 3.9%. This clearly indicates that the influential users identified from ARL are among the top ranked users being retweeted in #Bataclan. Regarding the statistical results of the AM, the results of GA and TS record values of 2%, 1% and 1.5%, SS results record 17.2%, 23% and 9.9% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are remarkably active on #Bataclan, moreover, they are among the users posting original tweets as indicated in SS results. Regarding the results of the TM, the results of BC record values of 38.2%, 27% and 36.4%, CC results record 38.7%, 31% and 30.3% and JWI results record 52.6%, 55% and 37.2% for ranked mean, median and standard deviation respectively. These results indicate that the influential users identified from ARL are not among the central and the socially connected users in #Bataclan. Among the distribution of mean values represented in Figures 3 to 5, we can clearly notice that in #ParisAttacks and #Bataclan, the mean values of UBR and the AM are between 2% and 17.2%. This indicates that the influential users in #Paris Attacks which are identified as influential using ARL are among the top active users and are among the top users being retweeted. Concerning #Paris, the UBR, GA and TS record approximately similar values and these values are within the top 25% which means that the identified influential users are active on #Paris, but the value of SS measure is relatively high compared to GA and SS and exceed the top ranked 25%. This designates that these influential users are not posting original tweets however, they are retweeting instead.

Table 10: Descriptive statistics of rating influential users from ARL with influence measures in #Bataclan

#Bataclan			
Influential users from ARL with Influence Measures	Mean	Median	Std.
Users from ARL with CPF	49.8	46	34.2
Users from ARL with UBR	3.3	1	3.9
Users from ARL with GA	2	1	1.5
Users from ARL with TS	2	1	1.5
Users from ARL with SS	17.2	23	9.9
Users from ARL with BC	38.2	27	36.4
Users from ARL with CC	38.7	31	30.3
Users from ARL with JWI	52.6	55	37.2

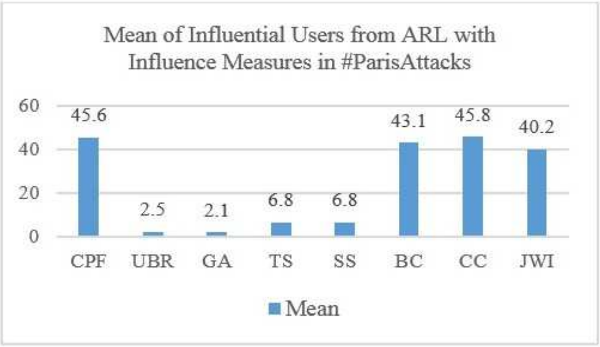

Fig. 3: Distribution of mean between influential users from ARL and influence measures in #ParisAttacks

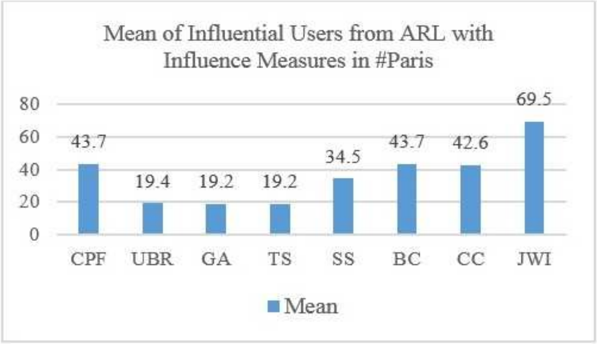

Fig. 4: Distribution of Mean between influential users from ARL and Influence Measures in #Paris

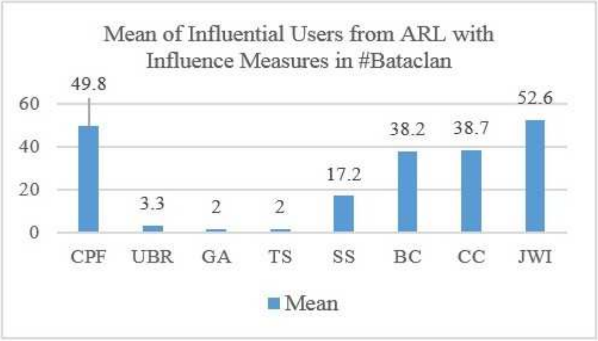

Fig. 5: Distribution of Mean between influential users from ARL and Influence Measures in #Bataclan

9.2 Findings

The findings from the social network analysis perspective reveal that the activity levels of users do not automatically influence the followers of these users to post on the same hashtags. As to TM, the JWI measure is a good indicator for user influence on other users posting on the same hashtag. Compared to the other two centrality measures, JWI results are valid. This means that the level of user social connectedness is a better indicator of the influence of the user than the other centrality measures. Furthermore, the name of the hashtag affects the level of the TM as noticed in Table 4. More familiar names of hashtags, such as #Paris and *#ParisAttacks*, has higher levels of similarity between the top rankings of CPF and TM. Also, the results reveal that in general hashtags, such as #Paris and #ParisAttacks, the user social connectedness, measured by JWI, is more influential on user's followers. While in focused hashtags, such as #Bataclan, JWI is not considered a good indicator as indicated in the results presented in Table 4 and Table 6, because as the social connectedness of users increases, the users are being retweeted more but are not influencing their followers to post on the same hashtags. Furthermore, the intersection results of SS and UBR for the different hashtags under study reveal that users who are tweeting in #Bataclan are being retweeted more than the users in #Paris and #ParisAttacks. This indicates that in focused hashtags, the content of the tweet itself influences the retweeting rate more than the level of users activity does.

Concerning results of the DM perspective, also interesting findings are revealed which reinforce the findings from the SNA perspective. From the ARL results, it is noticed that the users in each of the three hashtags tend to follow each other due to the high values of lift measure. Moreover, the identified influential users are not influencing their followers to post on same hashtags. In addition, the identified influential users are not socially connected neither central however they are active in all the hashtags under study. Also, in #ParisAttacks and #Bataclan, the identified influential users are not posting original tweets, however, they are retweeting.

Comparing the results of the two perspectives, it is noticed that there is an agreement between the findings from both perspectives. For example, the identified influential users from ARL method are noticeably active in the three hashtags under study, however, their followers aren't influenced by them to post on the same hashtags. Moreover, in #ParisAttacks and #Bataclan, all the identified influential users are clearly active, among the users being retweeted and posting original tweets. Whereas in #Paris, although identified influential users are also active and among the users being retweeted, but these users are not posting original tweets, they are retweeting instead.

10 Conclusion and Future Work

An analytical study has been conducted on a dataset collected from Twitter. The dataset consist of one of the top trendy hashtags between December 2015 and March 2016. The aim of this study is to identify influential users active on a specific hashtag. Two different perspectives are followed in order to identify influential users. In the first perspective, the SNA perspective, the influence measures are identified and their values are calculated. In the second perspective, the DM perspective, the ARL technique is adopted in order to predict influential users. The findings from the two perspectives are interesting revealing the importance of identifying influential users using ARL and validating the identified influential users using the Influence Measures. For future work, it would be very interesting to increase the size of the data set. Another interesting research area is to perform similar analysis on other hashtags of different type. Moreover, different categories of hashtags could be taken into consideration in order to test if determining influential users from different categories would yield different results.

References

[1] Z. Tufekci, Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls, Proceedings of the 8th International Conference on Weblogs and Social Media, pp. 505-514 (2014).

[2] C. Lee, H. Kwak, H. Park and S. Moon, What is Twitter, a social network or a news media?, Proceedings of the 19th international conference on World wide web, pp. 591-600 (2010).

[3] H. Efstathiades, D. Antoniades and G. Pallis, Online social network evolution: Revisiting the Twitter graph, Proceedings of the 2016 IEEE International Conference on Big Data, pp. 626-63 (2016).

[4] A. Patil, J. Liu and J. Gao, Predicting group stability in online social networks, Proceedings of the 22nd International Conference on World Wide Web, pp. 1021-1030 (2013).

[5] P. Brdka, S. Saganowski and P. Kazienko, GED: the method for group evolution discovery in social networks, Journal of Social Network Analysis and Mining, Vol. 3, No. 1, pp. 1-14 (2013).

[6] P. Brdka, P. Kazienko and B. Koloszczyk, Predicting Group Evolution in the Social Network, Proceedings of the International Conference on Social Informatics, pp. 54-67 (2012).

[7] S. Jamali and H. Rangwala, Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis, Proceedings of the 2009 International Conference on Web Information Systems and Mining, pp. 32-38 (2009).

[8] M. A. N. Hakim and M. L. Khodra, Predicting information cascade on Twitter using support vector regression, Proceedings of International Conference on Data and Software Engineering, pp. 1-6 (2014).

[9] M. Ilyas and H. Radha, Identifying Influential Nodes in Online Social Networks Using Principal Component Centrality, Proceedings of 2011 IEEE International Conference on Communications, pp. 1-5 (2011).

[10] B. Hajian and T. White, Modelling influence in a social network: Metrics and evaluation, Proceedings of 2011 IEEE Third International Conference on Social Computing, pp. 497-500 (2011).

[11] X. Jin and Y. Wang, Research on social network structure and public opinions dissemination of micro-blog based on complex network analysis, Journal of Network, Vol. 3, No. 1, pp. 1543-1550 (2013).

[12] A. Langville and C. Meyer, Deeper Inside PageRank, Journal of Internet Mathematics, Vol. 3, No. 1, pp. 335-380 (2003).

[13] Y.-R. Lin, D. Margolin, B. Keegan and D. Lazer, Voices of victory: a computational focus group framework for tracking opinion shift in real time, Proceedings of the 22nd international conference on World Wide Web, pp. 737-748 (2013).

[14] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy, Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Vol. 10, pp. 10-17 (2010).

[15] F. Erlandsson, A. Borg, H. Johnson and P. Brdka, Predicting User Participation in Social Media, Proceedings of the International Conference and School on Network Science, pp. 126-135 (2016).

[16] ”Hashtagify: Find, Analyse, Amplify,” CyBranding Ltd., 2011. [Online]. Available: https://hashtagify.me/explorer/. [Accessed 1 March 2015].

[17] M. Hawksey, ”The musing of Martin Hawksey (EdTech Explorer),” WordPress and Stargazer, 2011. [Online]. Available: https://mashe.hawksey.info/. [Accessed 10 November 2015].

[18] S. Niwattanakul, J. Singthongchai, E. Naenudorn and S. Wanapu, Using of Jaccard Coefficient for Keywords Similarity, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, No. 6, pp. 380-384 (2013).

[19] M. Bastian, S. Heymann and M. Jacomy, Gephi: An Open Source Software for Exploring and Manipulating Network, Proceedings of the Third International Conference on Weblogs and Social Media, Vol. 2009, pp. 361-362 (2009).

[20] L. I. Abu Daher, R. N. Zantout, I. T. Elkabani and K. Almustafa , Evolution of Hashtags on Twitter: A Case Study from Events Groups, Proceedings of the 5th symposium on Data Mining Applications, pp. 181-194 (2018).

[21] F. Riquelme and P. Gonzlez-Cantergiani, Measuring user influence on Twitter: a survey, Journal of Information Processing & Management, Vol. 52, No. 5, pp. 949-975 (2016).

[22] A. Pal and S. Counts, Identifying topical authorities in microblogs, Proceedings of the fourth ACM international conference on Web search and data mining, pp. 45-54 (2011).

[23] T. A. Kumbhare and . S. V. Chobe, An Overview of Association Rule Mining Algorithms, International Journal of Computer Science and Information Technologies, Vol. 5, No. 1, pp. 927-930 (2014).

[24] F. Erlandsson, P. Brdka, A. Borg and H. Johnson, Finding influential users in social media using association rule learning, Journal of Entropy, Vol. 18, No. 5, pp. 164-179 (2016).

[25] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the 20th International Conference on Very Large Data Bases, Vol. 1215, pp. 487-499 (1994).

[26] H. Tohidi and H. Ibrahim, A Frequent Pattern Mining Algorithm Based on FP-growth without Generating Tree, Proceedings of Knowledge Management 5th International Conference. Vol. 34, pp. 723-728 (2010).

Layal Abu Daher has obtained her Bachelor degree in Information Systems in 2007 and her Master degree in the same major in 2011 from Beirut Arab University, Lebanon. She has been a teaching assistant and a lecturer at Beirut Arab University since 2009 in addition to her administrative position at the Faculty of Science. Currently, Layal is a PhD candidate in Information Technology in the Department of Mathematics and Computer Science at Beirut Arab University, Lebanon. Her dissertation research is particularly in the field of Social Network Analysis and Mining.

Islam ElKabani received his Ph.D. in Computer Science from New Mexico State University (NMSU), USA in 2007. He worked as a Teaching and Research Assistant during his graduate studies at NMSU. Between 2007 and 2009, he worked as an Assistant Professor of Computer Science and an Executive Director of the Computer Center of the Faculty of Science at Alexandria University, Egypt. He has been a faculty member in the Computer Science Department at Beirut Arab University since September 2009. His research interests include Knowledge Representation, Answer Set Programming, Social Networks Analysis and Mining, Assistive Technologies, Natural Language Processing and Data Mining.

Rached Zantout received his B.E. from The American University of Beirut, Lebanon in 1988, his MSc from the University of Florida in 1990 and Ph.D. from the Ohio State University in 1994, all degrees being in Electrical Engineering. He was a Research Associate and Teaching Associate for most of his graduate studies. Directly after finishing his PhD he joined Scriptel Corporation and worked on several R&D projects to develop a new generation of graphic input devices. Between 10/1995 and 8/2000 Dr. Zantout was an Assistant Professor at King Saud University in Riyadh (Saudi Arabia). Then Dr. Zantout moved to Lebanon and taught as an Assistant Professor at the University of Balamand for the period between 9/2000 and 9/2002. He also worked as a part-time faculty members at reputed Lebanese universities like the American University of Beirut, Lebanese American University and Beirut Arab University. Between 9/2003 and 8/2009 Dr. Zantout was at the Hariri Canadian University where he became an Associate professor at the Electrical and Computer Engineering Department. Between 9/2009 and 9/2012 Dr. Zantout was an Associate professor at the College of Computer and Information Sciences at Prince Sultan University, Riyadh, Saudi Arabia. Between 9/2012 and 9/2014, Dr. Zantout was an Associate Professor at the Mathematics and Computer Science Department of the Faculty of Science at Beirut Arab University, Lebanon. Currently Dr. Zantout is Associate Professor at the Electrical and Computer Engineering Department of the College of Engineering at Rafik Hariri University. Dr. Zantout?s research interests cover Robotics, Artificial Intelligence, and Natural Language Processing. He currently works on developing components for Machine Translation and Natural Language Processing with a special focus on tools related to the Arabic Language. He also has active research in the area of autonomous robot navigation, Computer Vision, Digital Image Processing and Embedded Systems Design.