Simmelian ties on Twitter: empirical analysis and prediction

Isa Inuwa-Dutse, Mark Liptrott, Yannis Korkontzelos
Edge Hill Univeristy, United Kingdom
email: {*dutsei, Mark.Liptrott, Yannis.Korkontzelos*}*@edgehill.ac.uk*

***Abstract*— Social networks such as Twitter and Facebook come in various forms depending on the cohesiveness and size – from the most intimate to tenuous relationships. In the context of Twitter, the flexibility of establishing connections, such as a directed tie like following, enables the proliferation of tenuous relationships. This study observes that the implication of such flexibility poses challenges to data mining tasks, such as detection of socially cohesive groups, or content veracity. A small group of interconnected users or *Simmelian ties* are more intimate with a high degree of familiarity due to strong social cohesion. Such groups are considered homogeneous for many socio-demographic, behavioural, and intrapersonal characteristics. In the context of content veracity, anecdotal and cognitive evidence suggests that users are more likely to believe information shared by closely related individuals. Thus, the study is based on the premise that by recognising users who reciprocate friendships, some of the challenges will be mitigated. However, in social platforms such as Twitter, where flexible and transitory connections are prevalent, it is challenging to identify *Simmelian ties*.**

In this study, we present an empirical analysis of datasets consisting of 9300 Simmelian ties retrieved from over 30m Twitter accounts. Noting the challenges in identifying reciprocal relationships on a large scale, we propose a useful prediction model. As a result, the detection of socially cohesive communities is enhanced, thus providing a valuable analysis tool and strengthening the validity of online content. To evaluate the efficacy of the approach, we apply two state-of-the-art community detection algorithms on different datasets and achieve promising results. We further describe how to enhance content veracity and information diffusion by leveraging Simmelian connections. To the best of our knowledge, this study provides the first large scale dataset of *Simmelian ties* on Twitter.

***Index Terms* – Social networks, transitivity, Simmelian ties, clustering, Twitter**

I. INTRODUCTION

Humans are capable of attaching names to about 2000 faces but have a cognitive group size of only about 150, i.e. actively maintain social relationships. This limitation will be more pronounced in platforms such as Twitter where an average of 100m daily users contribute about 500m content messages[1], Twitter makes it difficult to keep track of socially cohesive groups. We argue that this challenge promotes the spread of irrelevant content and difficulty in the detection of local communities. The importance of a small group of users with a positive relationship has been recognised as a critical feature in the structural analysis of networks [13]. A small community of users (of approximately five members) are more intimate, with a high degree of familiarity due to strong social cohesion [11]. Focusing on smaller groups of users with reciprocal ties on Twitter is more helpful in discovering aspects of a *personal network* which is homogeneous for many socio-demographic, behavioural, and intrapersonal characteristics [28]. In social network analysis, the ability for a user to maintain a cohesive social relationship is impaired as the user's network size grows. This inverse relationship between *social cohesion* and *network size* is linked to the cognitive capability of the human brain. To motivate the approach, consider Fig. 1, which shows a summary of Dunbar's classification of social relationships as a function of *closeness* and *size*. Many theories and studies,

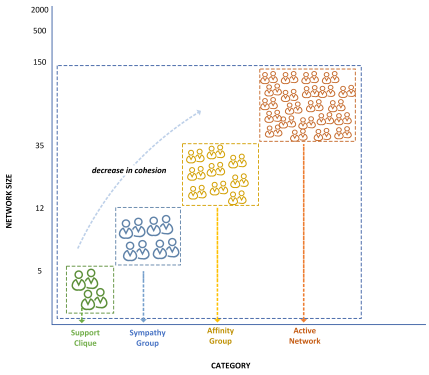

Fig. 1. Classification of social groups and the degree of cohesiveness. The smaller the size, the stronger the cohesion. Most connections on Twitter can be regarded as outside the active zone in the model. This means that Twitter users are expected to be less cohesive and large amounts of its content is expected to be irrelevant.

especially in social science, have analysed the relevance of socially cohesive groups, which constitute a community. A local community in a network is a crucial organising principle, especially in a vast network, and enables a better understanding of the structure and function of networks [32], [36], [38]. Moreover, members of a community are closely related and share many vital features that could be leveraged in analysis involving Twitter. For instance, anecdotal and cognitive evidence suggests that users are more likely to believe information from closely related individuals [7]. According to the theory of *cognitive balance*, if strong ties exist among three users, anything short of positive relationship would lead to *psychological strain* and would be avoided [30]. A better understanding among users in a social network is crucial in maintaining a civilised social space. However, as pointed earlier, the eccentricity of connections on Twitter make it difficult to establish reciprocal ties that could lead to *Simmelian* or *transitive ties*. We speculate

[1]See `www.omnicoreagency.com/twitter-statistics`

that this limitation contributes to identifying many socially unrelated users and encourage the propagation of spurious content. A *Simmelian tie* [35] is referred to as a strong social relationship within *three-person groups*. Simmelian tie originates from the work of [35] and defines how various social phenomena can be analysed in terms of relationships within three-person groups. The concept of *Simmelian tie* is similar to *transitivity*[2], a social preference to be friends with a *friend-of-a-friend* [36]. A relationship relation $\succ$, over a set $\{a,b,c,.....k\} \in D$ is *transitive*:

$$\text{iff } a \succ b \text{ and } b \succ c, \text{ then } a \succ c \;\; \forall a,b,c \in D$$

There is a little attempt to utilise *reciprocal ties* in analysis involving Twitter, which will ultimately lead to identifying transitive ties. Previous studies [3], [8], [21], [37] examine reciprocity for various tasks which are either based on *directed sets of nodes* or *textual content*, which do not convey the full meaning of reciprocity in the absolute term. A directed tie is peculiar to Twitter since, in other platforms, such as Facebook, an automatic reciprocal relationship is established once a friend request is accepted. In this study, we investigate the manifestation of *Simmelian ties* and how it could be leveraged in useful tasks such as clustering and mitigation of the spread of spurious online content. We treat a *Simmelian relationship* or a set of *transitive users* as a facilitator of social cohesion based on the premise: *if we can understand the underlying mechanisms to predict transitivity on Twitter, tasks such as cohesive clustering and content validation could be greatly enhanced*. Consequently, we put forward the following questions for investigation:

1) How can we identify and quantify the proportion of *Simmelian ties*?
2) How can we infer the latent variables, i.e. *reciprocity effect*, making it possible for two or more users to establish reciprocal relations?
3) How can we develop a *Simmelian relationship* prediction model and quantify the uncertainties surrounding the predictions?

To address these questions, we collected and analysed 9300 user profiles consisting of many reciprocated ties (*pairwise* and *transitive* ties, see Table II) gathered from over 30m Twitter accounts. We experiment with three different datasets consisting of undirected and directed ties and propose a model to predict the formation of a tie between any pair of users. We then explore ways to fully harness reciprocal ties towards enhancing community detection and content integrity. Through our study, we contribute the following:

- A large-scale empirical analysis of *Simmelian ties*. We describe how users in a *Simmelian triad* act as network bridges.
- A *Simmelian tie* prediction model that circumnavigates the difficult and time-consuming approach of finding *state-type ties* (see Section II). Using data from users with reciprocal relationships ensures efficiency and helps to mitigate the *curse of dimensionality* that could result from manual profiling of *Simmelian ties*.
- A thorough description of the applicability of *Simmelian ties* in enhancing clustering, information diffusion and as an effective means to reach out to socially cohesive groups of users.
- To the best of our knowledge, this study presents the first large scale empirical analysis of Simmelian ties on Twitter and the research data[3] is made available to support further research.

The remaining of the paper is structured as follows: we present the background and related word in section II and section III describes the proposed prediction framework. The discussion and conclusion are provided in sections IV and V, respectively. Table I shows a summary of the notations utilised in the study.

TABLE I
NOTATIONS AND DESCRIPTIONS USED IN THE STUDY

Notation	Description
D and $\hat{D}$	observed and synthetic data
θ	vector of unknown parameters, e.g. μ and σ
$p(\theta)$	prior distribution
$p(D\|\theta)$	likelihood function
$p(\theta\|D)$	posterior distribution
$M(\theta)$	generative model as a function of θ
β_{ui}	mean reciprocity among users
γc_{ui}	mean reciprocity between users' categories
ε_{ui}	error term in log-linear model y_i
χ_s	set of features inducing reciprocity
$a \succ b$	a binary relation between a and b
m	set of all followers of a user
τ	ratio of corresponding attributes with values $\in [0.75, 1.25]$
κ	set of reciprocal ties
y_i	log-linear model

II. NETWORK COMMUNITIES AND TIE FORMATION

A local community in a network is a crucial organising principle, especially in an extensive network and enables a better understanding of the structure and function of networks [32], [36], [38]. The structure and properties of various networks have been examined in the past [2], [12], [34], [36]. While many properties are shared across various networks, social networks exhibit different properties [32], which can be attributed to the methodological point of view. Social network theorists hold two methodological positions in investigating social relationships: *realist* and *nominalist*. The *realist* proceeds with a preconceived notion of the existence of relationships in a network which need to be discovered, whereas the *nominalist* relies on the questions asked by the investigator [22]. Moreover, a social tie can be an *event-type tie* or a *state-type tie*. An *event-type tie* is transitory and often results in socially distant members. With respect to Twitter, an *event-type tie* consists of subscription to *trending hashtags or retweeting popular users*; see Fig. 2. On

[2]transitive and Simmelian ties are synonymous in this study

[3]see https://github.com/ijdutse/simmelian_ties_on_Twitter for the dataset in accordance with Twitter sharing policy

the other hand, a *state-type tie* is based on static or structural connectivity between users, which suggests a certain degree of familiarity and trust [5]. We argue that the prevalence of directed connections on Twitter promotes the spread of irrelevant content and difficulty in the detection of local communities, among other challenges. For network analysis,

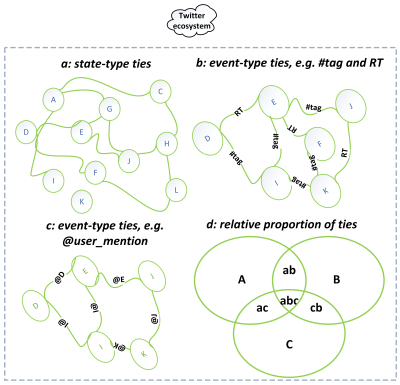

Fig. 2. The topological structure of Twitter allows for many forms of *event-type tie* to be formed ((a),(b) and (c)). Users openly connect with one another (a) – unidirectional or directed (friend or follower), bidirectional or undirected (both friend and follower) or (b – c) indirect or none (based on *transitory events* such as *retweets, mentions* or *likes*). These flexible connections also contribute to the proliferation of spurious content and (d) limit the number of cohesive social groups, as illustrated by the intersection region in (d).

transitivity is a vital feature of a network [36] that enables the formation of cohesive communities [15]. In the context of Twitter, a *Simmelian tie* refers to a set of transitive users connected via *state-type ties*. As illustrated in Fig. 2, the prevalence of transitory connections on Twitter makes it challenging to identify reciprocal ties based on *state-type* relationship. As a result, the problem of community detection on Twitter is mostly centred around the directed form of connections, i.e. *event-type ties* and adopt the *realist* approach. While this is valid in many networks, such an approach could lead to many unrelated sets of users.

A. Tie Formation

Network community has been considered as the basis for tie formation among users, where preference over a community (*shared community*) favours friendship between two users [25], [26], [38]. Users with few shared communities are more likely to connect than users with an uneven proportion of shared communities. For instance, if user v_1 is involved in 5 communities ($5c$); v_2 in $2c$; v_3 in $2c$ and a community is shared among all the users, then v_2 and v_3 are more likely to be friends than v_1 and v_2 or v_1 and v_3 [26]. This *community-centric* approach applies where a community is explicit in the network. In Twitter, communities could be formed based on many factors, as described in Fig.2, resulting in data far from adequate for cohesive community detection due to the prevalence of *event-type ties*. This study is interested in revealing factors that influence reciprocal ties that could ultimately result in a cohesive community on Twitter. We define the probability of tie formation from the

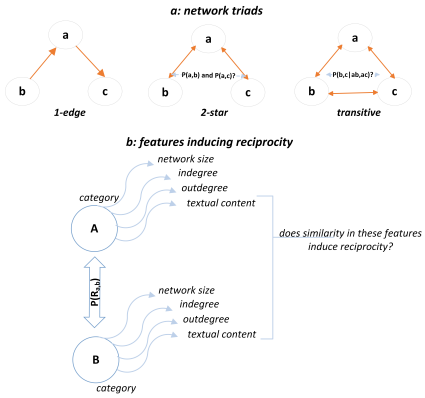

Fig. 3. (a) Possible triads in a network and degree of social ties. (b) Example of dyads with the corresponding network of users and relevant features responsible for the formation of a tie between users on Twitter. For each of these attributes, we assign a probability score to discover the inter-dependencies between the features in enabling reciprocal relationships. The union of the network of A and B is given by $A \cup B = a_1, a_2, a_3, ..., a_n, ..., b_1, b_2, b_3, ..., b_n$.

users perspective, i.e. in a *user-centric* manner, and propose a tie prediction model informed by the empirical analysis of reciprocal data.

III. Prediction Framework

In this section, we formulate the problem and describe our approach to solving it. We explain the *research data*, *meta-analysis*, *experimentation* and *evaluation* of the proposed model.

A. Problem Formulation

Considering Fig. 3(b), we are interested in understanding the link between the factors shown that are responsible for reciprocal ties among the users A, B, and C. A relation $\succ$ over a set D is *reciprocal* iff $a \succ b, \forall a, b \in D$. Similarly, a binary relation over a set D is *transitive* iff $\forall a, b, c \in D$, if $a \succ b$ and $b \succ c$ then $a \succ c$. Thus, given a set of users $a, b, c, ..., n \in V$ and a set of edges $e_1, e_2, ...e_n \in E \;\; V, E \in D$, the goal is to find the likelihood of *reciprocity* $p(R_{a,b})$ between any pair of users that could lead to a *Simmelian tie*.

B. Dataset

The study data consists of tweets collected from Twitter using a collection crawler that returns information about *reciprocated ties* and *unreciprocated ties* from each user's network. The collection begins with a set of users (*seed users*[4] from *verified* and *unverified* account categories. The dataset in [16] have been filtered to get rid of irrelevant content, and was used in earlier studies [17], [18]. Users on Twitter are broadly classified as *verified* or *unverified*. The *verified* refers to genuine users whose accounts have been authenticated or verified by Twitter. An *unverified* account is the one not authenticated by Twitter. Account verification can is viewed as an attempt to prevent fake accounts using

[4]We begin with 4022 genuine users obtainable from [16].

the name of politicians, celebrities, or popular individuals. For each user a consisting of m followers, we search and return the user if a reciprocal tie, $\succ$, exists between a and $b \in m$, $a \succ b = 1$ otherwise $a \succ b = 0$, i.e. $\exists b \in m : a \cap b = 1$. We denote the set of $a \succ b = 1$ by κ where $\kappa \in m$. From Fig. 2(d), if a commonality exists between the users a, b, c the search stops and transitive users are found. Table II shows a summary of the datasets.

TABLE II

DATASET SUMMARY. C: CATEGORY; S: SEED SIZE; V: VISITED USERS; P: PAIRWISE TIES; T: TRANSITIVE TIES; D: SEARCH DURATION

C	S	V	P	T	D (min.)
1:verified	1,000	1,832,630	708	–	1,122
2:verified	1,990	3,893,075	2,155	–	2,247
3:verified	6,803	14,413,641	1,317	541	7,965
1:unverified	1,000	1,793,806	640	–	2,162
2:unverified	2,023	13,409,661	1,834	–	4,084
3:unverified	7,121	32,065,133	2,150	347	13,071
ego-Twitter	81,306	–	–	–	–

In addition to the empirical data collected for the study, we use a benchmark data *(ego-Twitter)* dataset[5]. The dataset consists of directed ties only; hence, its applicability in this study is limited. We utilise it for prediction and comparison.

C. Meta-analysis

The rationale behind the *meta-analysis* is to identify user's attributes with strong correlation with *reciprocity effect* (see Fig. 3(b)). We apply a pragmatic approach to examine the distribution of ties and compute relevant metrics in the collected data.

1) Network topology: Fig. 4 shows the *empirical cumulative distribution function (ECDF)* of relevant metrics across user categories in the dataset. In the Fig. 4, there is a higher proportion of reciprocal ties in the *unverified* users category, and a plausible reason for the low proportion of reciprocal relationships in the *verified* users can be likened

[5]Obtained from Stanford public data repository [24].

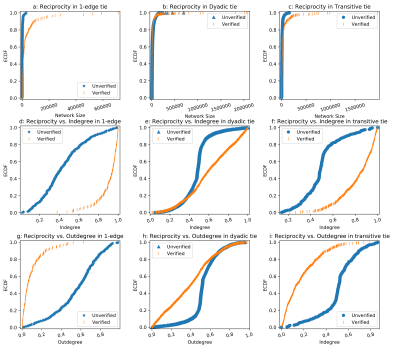

Fig. 4. The *ECDF* of various types of connections and network sizes in the data. The network neighbours of *verified* users are higher, but the *unverified* counterparts show a higher proportion of reciprocal ties. The relatively high proportion of 1-edge in the network can be explained by many followers not being followed back on Twitter.

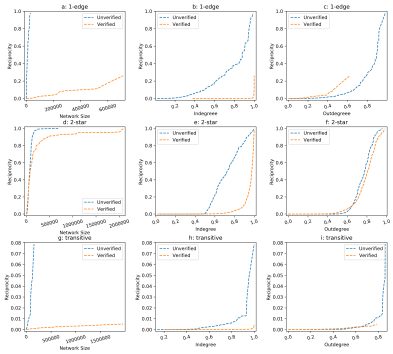

Fig. 5. The effect of user's attributes in enabling reciprocal ties. Both *indegree* and *outdegree* appear to be instrumental in enabling high degree of reciprocity. These provide useful insights about the effects of the user's attribute in influencing reciprocity which can be used to inform the prediction model.

to the reasons given in [8], that such users are authorities or institutions with independent sources of information outside the network. Based on this result, we can assume that the network size of a user is highly likely to grow if the user is *verified*, has *large followers*, but with decreasing *likelihood of reciprocity*. Similarly, there is a *high likelihood of reciprocity* if the user is *unverified* and has a relatively large network size. From Fig. 4, users in the *unverified category* are more likely to reciprocate a followership request, and users with the high number of network size (usually greater than 20k) have low proportion of reciprocated ties. Majority of reciprocated ties have network sizes below 20k. In Fig. 5, the proportion of reciprocity is higher for *unverified users*. We manually check some of the reciprocated accounts in the *verified* category and find that it is mostly reciprocating other verified. This reciprocal relationship can be attributed to trusting, i.e. the true identity of the users is known.

D. Bayesian Inference

Although the transitive relationship is rare on Twitter, finding it is significant since transitive users can be resourceful access-points to a more extensive network and credible information from the perspectives of the connected users. To improve the likelihood of users with reciprocal ties, we model the generative process of establishing a complementary relationship between users based on *Bayesian Inference*. The main goal is to investigate how the features identified in Fig. 3(b) affect a reciprocal relationship. The *reciprocity effect* sought to understand why some users have reciprocated ties, and some do not and how to predict the likelihood of reciprocal relations among users. We apply a simple *log-linear model* as a linear combination of the user's attributes to study the propensity of a user to reciprocate a tie (see 1). The *log-linear model* is commonly used in problems involving probabilistic prediction [4], [20].

$$y_i = \beta_{ui} + \gamma c_{ui} + \varepsilon_{ui} \tag{1}$$

In (1), $\beta_{ui}, \gamma c_{ui}$ and ε_{ui} denote the mean reciprocity among users, mean reciprocity between users' categories and error

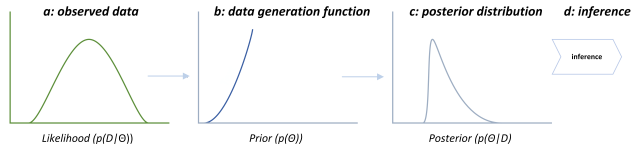

Fig. 6. A simple workflow in a hierarchical model for Bayesian inference. Basically, (a) the collected data/observations D (b) is assumed to be generated by a function of a set of unknown variables denoted by θ to (c) compute posterior distribution for (d) making an inference.

term respectively. The parameters in (1) are treated as random variables specified by probability distribution functions $p(\cdot)$ consisting of a range of values making it possible to define other statistical quantities such as mean μ. Table I describes the parameters and the distributions.

1) The Inference Workflow and Parameters: Fig. 6 shows the basic execution pipeline in the *Bayesian Inference* where the final hypothesis (d) i.e. inference is the estimated underlying probability correct for n trails in experiment j, e.g. θ_j. The *Prior* θ *and likelihood* $f(y|\theta,x)$ – represent set of variables that are likely to characterise the data informed by previous knowledge about the data. The assumption is θ_i comes from a probability distribution that describes the individual difference among users. The *posterior* $p(\theta|D)$ *or* $p(\theta|y,x)$ is given as a function of the *likelihood* and the *prior* which is simply the evidence in the data based on *Bayes' rule*. The rule entails updating beliefs about θ given the observed data D. Due to the small size of the data and time-consuming process of detecting a large amount of reciprocated ties on Twitter, the use of the Bayesian probabilistic approach makes it possible to simulate real data in a controlled setting before exposing the model to the small amount of the actual data. The essence of the *linear model* is to enable us to simulate the observed data D and generate a synthetic version $\hat{D}$ indistinguishable from the observed information D.

In Fig. 6, the data generation proceeds in the forward direction and the inference in the backward direction using the *linear model*. Finally, the inference (Fig. 6(d)) involves backtracking to determine the parameter that produced the observed data points. Many algorithms for inference, such as the maximum likelihood estimation [29] are used to estimate the parameter values that maximise the likelihood (given the observed data). We use *PYMC3 toolkit*[6] which incorporates all the required dependencies for our analysis. Fig. 7 and Fig. 8 shows data sampling and posterior distribution respectively. In Fig. 8, some of the users are measurable below the mean in the entire experiments, the *indegree*, for instance, is below the mean shown in the *meta-analysis* section. Although the mean is below the observed mean, it suggests that it is greater than chance, which will be useful in making credible assumptions about the data. For instance, we could quantify the uncertainties in the data when making predictions.

[6]A toolkit of probabilistic programming in Python [33]

E. Tie prediction

Inspired by the work of [1] in which attributes' similarities were applied for community detection tasks, the following section of the study describes how the probability of reciprocity $p(R_{a,b})$ or $p(R_{v_i,v_j})$ resulting from feature $f \in \chi_s$ is based on *attribute similarity* of users. As a form of *user-centric* approach, the prediction model is based on the following profile attributes, χ_s, *network size, indegree, outdegree, description information* and *tweet content*[7]. The *textual content* were not considered in this study. The *indegree (ind)* corresponds to the number of *followers* of a user; the *outdegree (out)* corresponds to the number *friends* or *followings* of a user; and the *category (cat)* denotes whether the account holder is *verified* or *unverified*. As illustrated in Fig. 3(b), these properties ($f \in \chi_s$) are shared by all users in V. Presumably, the decision to reciprocate is correlated with the idea of *homophily* in which we hypothesise that high similarity between users could be a reliable indicator of reciprocity. We ignore other latent factors that could induce reciprocity and assume that it is based on the available attributes identified in Fig. 3(b). Each attribute contributes to the overall decision based on its influence. Fig. 5 and Section III-D offer useful insights in this respect. The subset of the features[8] $\mathscr{X}_f$, for making the comparison consists of easily accessible attributes that enable a quick decision about reciprocity. Thus,

$$\{ind, out, cat\} \subset \mathscr{X}_f$$

It follows that, for a pair of nodes v_i, v_j, their corresponding features are given by

$$\mathscr{X}_{f_{v_i}} = \{ind_{v_i}, out_{v_i}, cat_{v_i}\}, \qquad \mathscr{X}_{f_{v_j}} = \{ind_{v_j}, out_{v_j}, cat_{v_j}\}$$

The ratio of the corresponding attributes, e.g. *ind or out*, between pairs is a real value quantity,

$$\frac{ind_{v_i}}{ind_{v_j}} \in \mathbb{R} \qquad \forall f \in \mathscr{X}_{f_{v_i,v_j}}$$

If the computation evaluates to a value within $[0.75, 1.25]$, the pairs are assumed to have similar attributes (1), or dissimilar attributes (0); this interval is to allow extra freedom for minor discrepancies between the corresponding features. For instance, if the ratio equals 1.0, the pairs have precisely similar attribute which is useful in analysing aspects of *homophily*. The binary values from the comparison of corresponding attributes or features are used to compute the overall similarity between pairs using *Jaccard Similarity Coefficient, J* (2):

$$J(\mathscr{X}_{f_{v_i}}, \mathscr{X}_{f_{v_j}}) = \frac{|\mathscr{X}_{f_{v_i}} \cap \mathscr{X}_{f_{v_j}}|}{|\mathscr{X}_{f_{v_i}} \cup \mathscr{X}_{f_{v_j}}|} \tag{2}$$

[7]textual features have been utilised in a related study by [18]

[8]For brevity, the features are trimmed, e.g. *indegree, outdegree, category* is trimmed to *ind, out, cat* respectively.

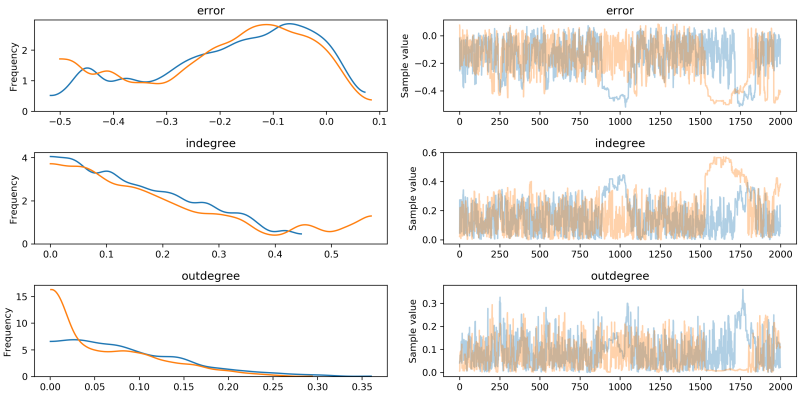

Fig. 7. Sampling results showing the *error term, indegree* and *outdegree*. Some of the samples are unstable, as evidenced by the perturbations in the results in the second column.

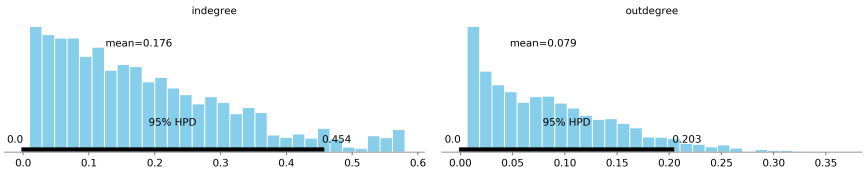

Fig. 8. Posterior distributions

a) Reciprocity & Constant Error Term: The response to a *friendship* request is either *yes* (reciprocate) or *no* (do not reciprocate), and is associated with a decision error, which is modelled using a *probabilistic preference model* or *response probability*. The response probability aims to capture various scenarios in which an actor is offered a set of features, and the decision process is associated with a constant probability of making an error in the choice [27]. The *probabilistic preference model* enables the mapping of each possible *response* into a probabilistic space, and utilised the *constant* or the *trembling hand* error, which assigns a constant value to a choice probability. The *error term* ζ, associated with each likelihood of reciprocity is based on the assumption that there is a $50-50$ chance of reciprocity or otherwise between any pairs in the network. Through the degree of similarities between the corresponding features, it is possible to improve the prediction by expressing the error term as a function of the similarity index $J(v_i, v_j)$, between pairs. Consequently, the prediction error ε_{v_i,v_j} (3), and the similarity index (2), are expressed in such a way that the prediction is within a practical significance range that closely match a realistic prediction using the following relation:

$$\varepsilon_{v_i,v_j} = \frac{1}{\zeta \times (1+\log(J(v_i,v_j)+\zeta))} \tag{3}$$

The symbol ζ corresponds to the *constant error term* and the final relation is given by:

$$p(R_{v_i,v_j}) = \frac{1}{1+\exp^{\varphi}} \tag{4}$$

where:

$$\varphi = -\log(\varepsilon_{v_i,v_j}+J(v_i,v_j)) \times (\varepsilon_{v_i,v_j}+J(v_i,v_j))$$

In this study, the constant error term $\zeta \geq 0.3$ and each item in the predicted ties, κ, satisfies (3) and the final relation given by (4). The value of ζ, is intuitively close to the reality of predicting ties on Twitter since many latent factors influence users' decision. The lower the error rate, the higher the chances of making a correct prediction, however, the constant error cannot be below 0.3 since it is highly unlikely to predict reciprocity with such precision noting the many factors that influence users' decision. With (4), it is possible to compute the probability of reciprocity between any pairs of nodes given their corresponding features. The prediction of reciprocity makes it possible to identify as many sets of nodes as possible with a high likelihood of establishing reciprocal ties, thus adding a layer of social cohesion tasks related to community detection. It follows that the likelihood of a reciprocal tie between any pair of users can be expressed as follows:

$$L(R_{v_i,v_j}) = 1 - \prod_{f \in \chi_s} (1 - p(R_{v_i,v_j})) \tag{5}$$

The relation $L(R_{v_i,v_j})$ (5) can be viewed as a generative process where $p(R_{v_i,v_j})$ is the marginal reciprocity effect of each feature $f \in \chi_s$ (see Table I).

IV. Utility of Simmelian Ties

This section discusses the significance of our findings and the applicability of those findings to *content veracity, information diffusion* and *community detection*.

A. Content Veracity and Diffusion

From the perspective of content integrity, a small group of users with reciprocal ties provides a useful means for analysing user groups with common online traits. According

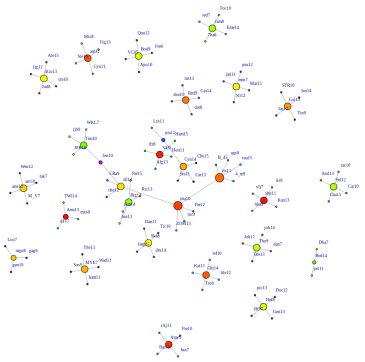

Fig. 9. An Example of users with reciprocated ties. For brevity, the actual network size of each user has been truncated. The size of each node in the figure reflects the user's network. The names of the users are anonymous to preserve identity. We retain only the first three letters of each screen-name and attach the length of the name as postfix.

to [15], the hypothesis that allows a strong reciprocal tie ($A \Longleftrightarrow B$) is given by *the stronger the tie between A and B, the larger the proportion of entities* (i.e. in S, see Fig. 3(b)) to whom they will both be tied (weak tie), i.e. connected by a weak tie or strong tie. In Fig. 3(b), there is less overlap in friendship circles ($a \cap b \in S$) if the tie between A and B is non-existent; intermediate if it is weak and most when it is strong. Similarly, if strong ties connect $A \leftrightarrow B$ and $A \leftrightarrow C$, both C and B, is similar to A, hence the likelihood of friendship increases once they met. It has been suggested that if strong ties exist among three users, anything short of positive relation will lead to a *psychological strain* [30] and increases the likelihood of losing a third-party relationship [6].

1) Information Diffusion: A characteristic feature of a network is the presence of a core-periphery pattern with a central group of closely related users. Usually, these users (acting as social bridges) are less connected to the core network and one another but play a significant role in connecting disparate parts of a community [6]. We refer to users with a high proportion of reciprocated ties as *hop-skippers*[9], see Fig. 9. Hop-skippers provide the means, i.e. the local information required in local community detection or cohesive community detection by their centrality. A user with many reciprocal ties would be a resourceful representation of users with strong social cohesion.

B. Community detection

The target of a clustering algorithm is to identify a high degree of similarity within a community of users using a scoring function that enables the grouping of objects according to the extent to which they are equivalent using a set of experimental procedures. Depending on the procedural approach, the definition of equivalence usually leads to different partitioning of a network. Network objects can be equivalent (1) if they have the same connection pattern to the same neighbours or (2) if they have the same or similar connection pattern to different neighbours [10]. In the context of this study, equivalence relates to the structural similarity of users on Twitter.

[9]We use the term *hop-skippers* (á la [19]) to denote users with large number of reciprocal ties in Twitter.

TABLE III

EXPERIMENT ON THREE DIFFERENT DATASETS FOR COMMUNITY DETECTION USING TWO DIFFERENT ALGORITHMS. G–N AND LP: GIRVAN–NEUMAN AND LABEL PROPAGATION RESPECTIVELY; #DC: NUMBER OF DETECTED COMMUNITIES

Dataset	G–N			LP		
	Metric		#DC	Metric		#DC
	Q	NMI		Q	NMI	
Ground-truth	.908	.794	308	.77	.602	1319
ego-Twitter	.334	.197	1431	.215	.131	2131
Predicted	**.473**	**.311**	1107	**.360**	**.267**	2071

1) Clustering Based on a Set of Reciprocal Ties: A local community is defined based on the local information about the network where its members have strong social cohesion. To demonstrate the relevance of *Simmelian ties* to a community detection task on Twitter, we utilised two state-of-the-art community detection algorithms: *Girvan-Newman (G-N)* [14] and *Label Propagation (LP)* [39]. The LP algorithm is an iterative clustering method suitable for use with unlabelled data that operates by turning it to a labelled data using an initial set of labelled data. The idea in the algorithm is to propagate the labelled information across the whole dataset.

2) Evaluation: We experiment on three different datasets using the algorithms. The ground-truth dataset achieves higher performance followed by the dataset with predicted ties and lastly the *ego-Twitter* (Table III). The high performance in the dataset with *predicted ties* attests to the relevance of reciprocated ties in enhancing clustering. For the evaluation, we aim to answer the following questions:

1) *can we effectively predict reciprocal ties that could lead to the detection of local communities in Twitter?*
2) *does the inclusion of hop-skippers in local community detection improve performance in comparison to a standard approach?*

The algorithms, as mentioned earlier, are applied to the data based on the experimental design and compare performance (Table III). The evaluation metrics consist of *Modularity measure* and *Normalised Mutual Information*. The *Modularity (Q)* proposed in [31] refers to the modularity metric that measures the strength of the identified communities, higher values are preferred (2) *Normalised Mutual Information (NMI)* [9] also evaluates the quality of clusters detected by various algorithms.

There is an active correspondence between the ability of a clustering algorithm to correctly identify groups and the *signal-to-noise-ratio* within the matrix of instances [23]. Many ideas can be explored to improve the detection of socially cohesive communities, e.g. using a robust similarity function to construct affinity matrix which can be used to analyse the *information content* in the clustering data. Based on the structural similarity and relevant heuristics, a more robust and effective clustering can be achieved.

V. CONCLUSION

Modern social media platforms enable the empirical quantification and evaluation of social relationships among users on an unprecedented scale. A *Simmelian triad* consists of a small cohesive group which reflects a personal network on Twitter, which is homogenous concerning many socio-demographic behavioural, and intra-personal characteristic [28]. This study is based on the assumption that a clustering method that recognises *Simmelian ties* offers a more transparent and cohesive representation of a community. However, *Simmelian ties* rarely occur on Twitter and they differ depending on the user category – *verified* or *unverified* among others. We applied a pragmatic approach to examining reciprocal ties – *pairwise* and *transitive* and conducted an empirical analysis to understand *Simmelian ties* on Twitter, where connections among users are porous. We analysed such ties and presented a formal prediction framework. Our findings suggest that users with *Simmelian ties* exhibit useful behaviours such as connecting large groups of users or acting as network bridges on Twitter. We demonstrated how *Simmelian ties* could be utilised in crucial tasks such as clustering and content integrity.

ACKNOWLEDGMENT

This research work is part of the CROSSMINER Project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under grant agreement No. 732223.

REFERENCES

[1] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *nature*, 466(7307):761, 2010.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Ego networks in twitter: an experimental analysis. In *2013 Proceedings IEEE INFOCOM*, pages 3459–3464. IEEE, 2013.

[4] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.

[5] Stephen P Borgatti and Daniel S Halgin. On network theory. *Organization science*, 22(5):1168–1181, 2011.

[6] Daniel J Brass. Men's and women's networks: A study of interaction patterns and influence in an organization. *Academy of Management journal*, 28(2):327–343, 1985.

[7] Kathleen Carley. A theory of group stability. *American sociological review*, pages 331–354, 1991.

[8] Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi, and Krishna Gummadi. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(4):991–998, 2012.

[9] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[10] Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj. Positional analyses of sociometric data. *Models and methods in social network analysis*, 77:77–96, 2005.

[11] Robin IM Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 6(5):178–190, 1998.

[12] Paul Erdös and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[13] Linton C Freeman. Some antecedents of social network analysis. *Connections*, 19(1):39–42, 1996.

[14] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[15] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[16] Isa Inuwa-Dutse, Mark Liptrott, and Ioannis Korkontzelos. Detection of spam-posting accounts on twitter. *Neurocomputing*, 315:496–511, 2018.

[17] Isa Inuwa-Dutse, Mark Liptrott, and Ioannis Korkontzelos. A deep semantic search method for random tweets. *Online Social Networks and Media*, 13:100046, 2019.

[18] Isa Inuwa-Dutse, Mark Liptrott, and Yannis Korkontzelos. Analysis and prediction of dyads in twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 303–311. Springer, 2019.

[19] Jane Jacobs. The death and life of great american cities. 1961. *New York: Vintage*, 1992.

[20] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[21] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[22] Edward O Laumann, Peter V Marsden, and David Prensky. The boundary specification problem in network analysis. *Research methods in social network analysis*, 61:87, 1989.

[23] Daniel John Lawson and Daniel Falush. Population identification using genetic data. *Annual review of genomics and human genetics*, 13:337–361, 2012.

[24] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[25] Hsin-Min Lu and Chien-Hua Lee. A twitter hashtag recommendation model that accommodates for temporal clustering effects. *IEEE Intelligent Systems*, 30(3):18–25, 2015.

[26] Chen Luo and Anshumali Shrivastava. Jaccard affiliation graph (jag) model for explaining overlapping community behaviors. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1–8. IEEE, 2018.

[27] AAJ Marley and M Regenwetter. Choice, preference, and utility: Probabilistic and deterministic representations. *New handbook of mathematical psychology*, 1:374–453, 2016.

[28] J Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[29] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[30] Theodore M Newcomb. The acquaintance process: Looking mainly backward. *Journal of Personality and Social Psychology*, 36(10):1075, 1978.

[31] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[32] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

[33] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[34] John Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[35] Georg Simmel et al. The stranger. *The Sociology of Georg Simmel*, 402:408, 1950.

[36] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440, 1998.

[37] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[38] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *2012 IEEE 12th International Conference on Data Mining*, pages 1170–1175. IEEE, 2012.

[39] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.