Interest Clustering Coefficient: a New Metric for Directed Networks like Twitter

Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, Stéphane Pérennes

Interest Clustering Coefficient: a New Metric for Directed Networks like Twitter*

Thibaud Trolliet[1], Nathann Cohen[2], Frédéric Giroire[2], Luc Hogie[1], and Stéphane Pérennes[2]

[1] INRIA Sophia-Antipolis, France
[2] Université Côte d'Azur/CNRS, France

Abstract. We study here the clustering of *directed* social graphs. The clustering coefficient has been introduced to capture the social phenomena that a friend of a friend tends to be my friend. This metric has been widely studied and has shown to be of great interest to describe the characteristics of a social graph. In fact, the clustering coefficient is adapted for a graph in which the links are undirected, such as friendship links (Facebook) or professional links (LinkedIn). For a graph in which links are directed from a source of information to a consumer of information, it is no longer adequate. We show that former studies have missed much of the information contained in the directed part of such graphs. We thus introduce a new metric to measure the clustering of a directed social graph with interest links, namely the interest clustering coefficient. We compute it (exactly and using sampling methods) on a very large social graph, a Twitter snapshot with 505 million users and 23 billion links. We additionally provide the values of the formerly introduced directed and undirected metrics, a first on such a large snapshot. We exhibit that the interest clustering coefficient is larger than classic directed clustering coefficients introduced in the literature. This shows the relevancy of the metric to capture the informational aspects of directed graphs.

Keywords: Complex networks, Clustering Coefficient, Directed networks, Social networks, Twitter.

1 Introduction

Networks appear in a large number of complex systems, whether they are social, biological, economical or technological. Examples include neuronal networks, the Internet, financial transactions, online social networks, ... Most "real-world" networks exhibit some properties that are not due to chance and that are really different from random networks or regular lattices. In this paper, we focus on

*This work has been supported by the French government through the UCA JEDI (ANR-15-IDEX-01), EUR DS4H (ANR-17-EURE-004) Investments in the Future projects, and ANR DIGRAPHS, by the SNIF project, and by Inria associated team EfDyNet. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

the study of the clustering coefficient of social networks. Nodes in a network tend to form highly connected neighborhoods. This tendency can be measured by the clustering coefficient. It is classically defined for undirected networks as three times the number of triangles divided by the number of open triangles (formed by two incident edges). This clustering coefficient had been computed in many social networks and had been observed as much higher than what randomness would give. Triangles thus are of crucial interest to understand "real-world" networks.

However, a large quantity of those networks are in fact directed (e.g. the web, online social networks like Instagram, financial transactions). It is for instance the case of Twitter, one of the largest and most influential social networks with 126 million daily active users [14]. In Twitter, a person can follow someone she is interested in; the resulting graph, where there is a link $u \to v$ if the account associated to the node u followed the account associated to the node v, is thus directed. In this study, we used as main dataset the snapshot of Twitter (TS in short) extracted by Gabielkov et al. as explained in [6] and made available by the authors. The TS has around 505 million nodes and 23 billion arcs, making it one of the biggest snapshots of a social network available today.

The classic definition of the clustering coefficient cannot be directly applied on directed graphs. This is why most of the studies computed it on the so-called *mutual graph*, as defined by Myers & al. in [11], i.e., on the subgraph built with only the bidirectional links. We call *mutual clustering coefficient (mcc for short)* the clustering coefficient associated with this graph. We computed this coefficient in the TS, using both exact and approximated methods. We find a value for the mcc of 10,7%. This is a high value, of the same order as those found in other web social networks.

However, this classical way to operate *leaves out 2/3 of the graph!* Indeed, we computed that the bidirectional edges only represents 35% of the edges of the TS. A way to avoid it is to consider all links as undirected and to compute the clustering coefficient of the obtained undirected graph. We refer to the corresponding computed clustering coefficient as the *undirected clustering coefficient (ucc for short)*. Such a computation in the TS gives a value of ucc of only 0.11%. This is way lower than what was found in most undirected social networks. It is thus a necessity to introduce specific clustering coefficients for the directed graphs. More generally, when analyzing any directed datasets, it is of crucial importance to take into account the information contained in its directed part in the most adequate way.

A first way to do that is to look at the different ways to form triangles with directed edges. Fagiolo computed the expected values of clustering coefficients considering directed triangles for random graphs in [5] and illustrated his method on empirical data on world-trade flows. There are two possible orientations of triangles: transitive and cyclic triangles, see Figures 1b and 1c. Each type of triangles corresponds to a directed clustering coefficient:

– the ***transitive clustering coefficient*** (***tcc*** in short), defined as:

$$tcc = \frac{\#\text{ transitive triangles}}{\#\text{ open transitive triangles}},$$

– the ***cyclic clustering coefficient*** (***ccc*** in short), defined as:

$$ccc = \frac{3 \cdot \#\text{ cyclic triangles}}{\#\text{ open transitive triangles}}.$$

We computed both coefficients for the snapshot, obtaining $tcc = 1.9\%$ and $ccc = 1.7\%$. However, note that a large part of the transitive and cyclic triangles comes from bidirectional triangles. When removing them, we arrive to values of $tcc = 0.51\%$ and $ccc = 0.24\%$.

We believe those metrics miss an essential aspect of the Twitter graph: while the clustering coefficient was defined to represent the social cliques between people, it is not adequate to capture the information aspect of Twitter, known to be both a social and information media [8, 11]. In this work, we go one step further in the way directed relationships are modeled. We argue that in directed networks, *the best way to define a relation or similarity between two individuals (Bob and Alice) is not always by a direct link, but by a common interest*, that is, two links towards the same node (e.g., Bob → Carol and Alice → Carol). Indeed, when discussing interests, consider two nodes having similar interests. Apart from being friends, these two nodes do not have any reason to be directly connected. However, they would tend to be connected to the same out-neighbors. We exploit this to study a new notion of connections in directed networks and the new naturally associated clustering coefficient, which we name *interest clustering coefficient, or* `icc` in short, and define as follows:

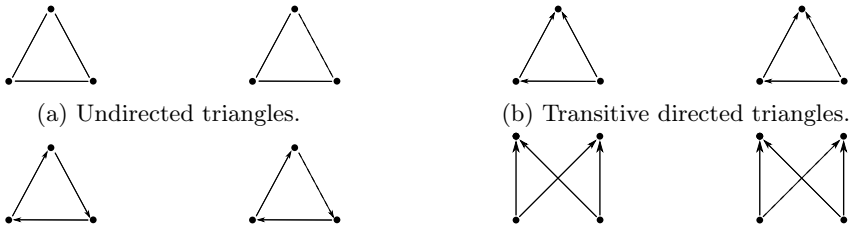
$$\texttt{icc} = \frac{4 \cdot \#\text{ K22s}}{\#\text{ open K22s}},$$

where a K22 is defined as a set of four nodes in which two of them follow the two others, and an open K22 is a K22 with a missing link, see Figure 1d. We computed the `icc` on the Twitter snapshot, obtaining `icc` = 3.6% (3.1% when removing the bidirectional structures). This value, an order of magnitude higher than the previous clustering coefficients computed on the non bidirectional directed graph, confirms the interest of this metric. If the clustering coefficient of triangles are good metrics to capture the social aspect of a graph, the interest clustering coefficient is a good metric to capture the informational aspect.

In summary, our contributions are the following:

– We define a new clustering coefficient for graphs with interest links.
– We succeeded in computing it, both exactly and using sampling methods, for a snapshot of Twitter with 505 million nodes and 23 billion edges.
– We additionally provide the values of the directed and undirected clustering coefficients previously defined in the literature. We believe this is the first time that such coefficients are computed exactly for a large *directed* online social network.

The paper is organized as follows. We first discuss related work in Section 2. In Section 3, we present the algorithms we used to compute the values of the interest clustering coefficient, both exactly and by sampling. We discuss the results on the clustering coefficients of Twitter in Section 4 .



(a) Undirected triangles. (b) Transitive directed triangles.

(c) Cyclic directed triangles. (d) K22s.

Fig. 1: Closed (left) and open (right) undirected and directed triangles and K22s.

2 Related Work

Clustering coefficient. The clustering coefficient shows that, when two people know each other, there is a high probability that those people have common friends. The clustering coefficient has numerous important applications, such as spam detection [4], link recommendation [15], information spread [7], etc. There are different definitions of the clustering coefficient. The *global clustering coefficient*, sometimes also called *transitivity*, was first introduced by Barrat and Weigt in [2]. It is defined as 3 times the number of triangles in the graph, divided by the number of connected triplets of vertices in the graph. Another definition was given by Watts and Strogatz [19] and is called the *local clustering coefficient*. It is defined as the mean over all nodes of the graph of the local clustering of each node, that is the probability that two random neighbors of the node are also connected together. We use the global clustering coefficient in this paper. The clustering coefficient has also been defined for weighted graphs [13].

Computations for social graphs. The undirected clustering coefficient of some social networks has been provided in the literature. It has been computed on very large snapshots for Facebook [17], Flickr, and YouTube [10]. The local clustering coefficient has also been studied in the undirected mutual graph of Twitter [11]. The undirected clustering coefficient is usually much higher in social networks than in random models.

Directed graphs. All these studies only consider the undirected clustering coefficient, even for directed graphs like Twitter. Fagiolo introduced definitions of directed clustering coefficients, that we named tcc and ccc [5], but those definitions had never been computed and discussed on large datasets to our knowledge, as we do in this paper. Moreover, we believe that these metrics are *not the most relevant ones for directed graphs with interest links.*

Computing substructures. Researchers studied methods to efficiently compute the number of triangles in a graph, as naive methods are computationally very expensive on large graphs. Two families of methods have been proposed: triangle exact counting or enumeration and estimations. In the first family, the fastest algorithm is due to Alon, Yuster, and Zwick [1] and runs in $O(m^{1.41})$, with m the number of edges. However, methods using matrix multiplication cannot be used for large graphs because of their memory requirements. In practice, enumeration methods are often used, see e.g., [9]. Methods to count rectangles and butterfly structures in undirected bipartite networks were also proposed in [18] and in [12]. In this paper, we propose an efficient enumeration algorithm to count the number of K22s and open K22s in a very large graph. We focused on the case in which only one adjacency can be stored, as this was our case for the TS. To the best of our knowledge, we are the first to consider this setting.

3 Computing Clustering Coefficients in Twitter

To compute the interest clustering coefficient and the triangle clustering coefficients of very large networks, we used two different methods presented here: an exact count and an estimation using sampling techniques, either with a Monte Carlo algorithm or with a sampling of the graph. As a typical example of a massive directed social network with interest links, we carried out the computations for a Twitter snapshot (TS in short) with 505 million nodes and 23 billion links, described in Report [16].

3.1 Exact Count

We computed the exact numbers of K22s and open K22s in the Twitter Snapshot. Recall that we are discussing a dataset with hundreds of million nodes and billions of arcs. Results are reported in Table 1 and discussed in section 4. We first present in this section the algorithms we use, and discuss their complexity.

In the rest of this paper, we call *top vertices* (resp. *bottom vertices*) of a K22 the vertices which are destinations (resp. sources) of the K22 edges. We call a *fork* a set of two edges of a K22 connected to the same vertex. We say that a *fork has top (or bottom) vertex x* if both edges are connected to x and x is a top (resp. bottom) vertex of the K22. The same terminology applies to open K22s.

Trivial algorithm. The trivial algorithm would consider all quadruplets of vertices with 2 upper vertices. Then, for each quadruplet, it would check the existence of a K22 and of open K22s. There are $\binom{4}{2}\binom{n}{4}$ such quadruplets. It thus gives a complexity of $O(n^4)$. This method can thus not be considered for the TS as it would perform 6.4×10^{33} iterations.

Improved algorithm. The practical complexity can be greatly improved by only considering *connected quadruplets*, and by mutualizing the computations of

the common neighbors of the in-neighbors of a vertex, as explained below. The pseudo-code is given in Algorithm 1.
The algorithm's main loop iterates on the vertices of the graph. For each vertex x, we consider its in-neighborhood $N^-(x)$. We then compute how many times a vertex w (with $w < x$ to avoid counting a K22 twice) appears in the out-neighborhoods of the vertices of $N^-(x)$. We denote it $\#occ(w)$. We use a table to store the value of $\#occ(w)$ in order to be able to do a single pass on each out-neighbor.
For a vertex w, any pair of its $\#occ(w)$ in-neighbors common with x forms a K22 with x and w as top vertices. There are hence $\binom{\#occ(w)}{2}$ K22s with x and w as top vertices. The number of K22s with x as a top vertex is then

$$\#K22(x) = \sum_{w|\#occ(w)\geq 2} \binom{\#occ(w)}{2}.$$

The number of open K22s with x as the top vertex is computed by noticing that, for any pair of vertices u and v of $N^-(x)$, we have $d^+(u) - 1 + d^+(v) - 1 - \mathbb{1}_{v\in N^+(u)} - \mathbb{1}_{u\in N^+(v)}$ open K22s containing this fork (ux, vx). We can count the number of open K22s with x as a top vertex, u as the bottom vertex of out-degree at least 2 (and thus another vertex v as the bottom vertex of out-degree at least 1). A vertex $u \in N^-(x)$ is thus in $(d^+(u) - 1 \sum_{v\in N^-(x)\setminus\{u\}} 1_{v\in N^+(u)})(d^-(x) - 1)$ such open K22s. The only subtlety is that we count the number of arcs, which are between two vertices of $N^-(x)$, during the loop on the out-neighborhoods of the vertices of $N^-(x)$. We note this number #internalArcs. We then have:

$$\#openK22(x) = \left(\sum_{u\in N^-(x)} (d^+(u) - 1)(d^-(x) - 1) \right) - \#internalArcs.$$

Lastly, the global number of K22s (resp. open K22s) in the digraph is just the sum of the number of K22s (resp. open K22s) with a vertex x as a top vertex, as, since we only consider K22s formed with a vertex w such that $x < w$, we only count each K22 once.

Complexity of the used algorithm. The complexity thus is $m + \sum_u d^+(u)(d^+(u) - 1)$, with m the number of edges. Indeed, each edge is only considered once as an in-arc and $d^+ - 1$ times as an out-arc. Note that, in the Twitter Snapshot, the sum of the squares of the degrees is equal to $8 \cdot 10^{13}$. The order of the number of iterations needed to compute the number of K22s was thus massively decreased from the 6.4×10^{33} iterations of the trivial algorithm.
For graphs following a power-law degree distribution with exponent between 2 and 3, we show in [16] that this gives a complexity between $O(m+n)$ and $O(n^2)$, to be compared to the one of the naive method $O(n^4)$.
Note that the number of undirected and directed triangles can be easily computed while counting the K22s, see [16].

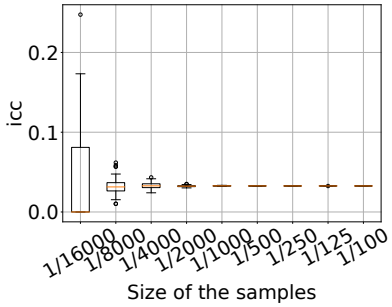
Algorithm 1 Enumeration of K22s and open K22s

Input: Digraph(V, A)
#occ=0 ▷ table
for $x \in V$ **do**
  $\#internalArcs \leftarrow 0$ ▷ We count the number of arcs internal to $N^-(x)$ as these arcs do not form open K22s
  for $v \in N^-(x)$ **do**
    #openK22s $+= (d^+(v) - 1)(d^-(x) - 1)$
    for $w \in N^+(v) \setminus \{x\}$ **do**
      #occ$[w]+ = 1$
      if $w \in N^-(x)$ **then** ▷ We use a second table to test that.
        $\#internalArcs+ = 1$
  for w with #occ$[w] \geq 2$ **do**
    #k22$+ = \binom{\#occ[w]}{2}$
  #openK22s $-= \#internalArcs$
  $\#occ \leftarrow 0$ ▷ Done with a double loop
icc $\leftarrow \frac{4\#K22}{\#openK22}$

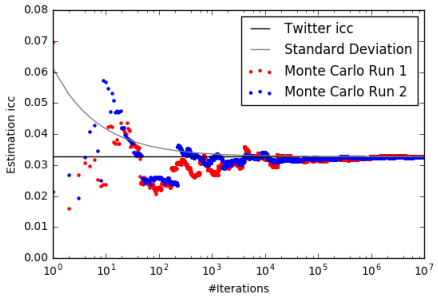
3.2 Approximate Counts

As discussed later in Section 4, the exact count of the number of K22s and open K22s in Twitter implies massive computations. This number can be estimated using Monte Carlo Method and/or computations on a sample of the graph. We discuss both methods below. One of our goals was to see how good computations made in the literature using smaller Twitter snapshots were.

Exact `icc` on Twitter Samples. We built samples of the TS to estimate the interest clustering coefficient. Several choices can be made to build the samples. To avoid missing nodes of high degrees (which would lead to a high variance), we sampled the arcs (and not the nodes). Given a sampling probability p, we keep an arc in the sample with probability p. We generated samples of different sizes corresponding to sampling probabilities from $p = 1/100$ to $p = 1/16000$.
Estimator of the number of K22 and open K22s. We use the classic estimator $X = \sum_{A \in \mathcal{A}} X_A$, where X_A is the random variable which is equal to 1 if all the arcs of pattern A are selected in the sample and 0 otherwise. Theoretical bounds for expectation and variance are given in [16]. The difficulty for the variance comes from the fact that two K22s can share a common link.
Results. We present in Figure 2a the results of the algorithm for different sample sizes, corresponding to sampling probabilities from $p = 1/100$ to p$=1/16,000$. For each sample size, we generated 30 samples. The distribution over the samples of the interest clustering coefficient is provided by a boxplot for each value of p. Note that a K22 of the TS appears in a sample with a probability of only p^4, and of p^3 for an open K22. The clustering coefficient of a sample is thus an estimate of $p \cdot$ `icc`.
We observe that, the clustering coefficient is well estimated using any sample for



(a) Estimation of the interest clustering coefficient for different sample sizes with Edge Sampling Method.



(b) Estimation of the interest clustering coefficient with Monte Carlo Method.

Fig. 2: Estimation of the interest clustering coefficient with Approximate Counts.

a sampling probability of 1/1000 or larger. Indeed, for this range of probabilities, the distribution over all samples is very concentrated and around the exact value of the `icc`. Note that, for $p = 1/1000$, a K22 is present in the sample with a probability of only 10^{-12}. The expectation of the number of nodes with an edge is only 23 million nodes (over 500 million) and the number of edges is also around 23 million. Thus, in the TS, a small sample (5% of the nodes and 0.1% of edges) allows to do an efficient estimation of the `icc`.
For smaller values of p, the variance increases. The median estimates well the `icc` for a range of p between 1/8000 and 1/1000, but samples of these sizes may have error of 100% of the value. Lastly, for $p = 1/16000$, most of the time there is no surviving K22 in the sample, leading to a value of zero for the `icc`. The `icc` thus cannot be estimated correctly.
In conclusion, a sample with sampling probability 1/1000 is enough to efficiently estimate the interest clustering coefficient, with a computation time of around 1 minute (instead of days for the whole TS) on a machine of the cluster.

Monte Carlo Method. The difficulty to estimate the clustering coefficients using Monte Carlo Method is that the probability to observe a (closed or open) K22 or a triangle is very small. In the case of triangles, this difficulty can be easily circumvented by knowing the node degrees. This allows to select an open triangle uniformly at random. In the case of K22s, this information is not sufficient to select an open K22 uniformly at random. In fact, achieving this goal is very costly. We thus present a *new method in which, by picking only forks* (as we do for triangles), we can compute the interest clustering coefficient. The idea is to select a vertex v as a root according to the square of its in-degree (as in the case of triangles), but without knowing its number of open K22s (first step). We then select two arcs u_1v and u_2v uniformly at random (second step). We then compute the number of K22s and open K22s with the selected fork (u_1v, u_2v) (third step). The formal justification of this new method is developed in [16]. We present here the results of the experiments.

	$\#closed$	$\#open$	cc
icc	$2.6 \cdot 10^{16}$	$3.1 \cdot 10^{18}$	3.3%
tcc	$2.5 \cdot 10^{12}$	$1.3 \cdot 10^{14}$	1.9%
ccc	$7.2 \cdot 10^{11}$	$1.3 \cdot 10^{14}$	1.7%
ucc	$6.2 \cdot 10^{11}$	$1.6 \cdot 10^{15}$	0.11%
mcc	$3.2 \cdot 10^{11}$	$8.9 \cdot 10^{12}$	10.7%

Table 1: Clustering coefficients in the TS. The first (resp. second) column represents the number of closed (resp. open) K22s or triangles in the Twitter Snapshot. Each line corresponds to a clustering coefficient defined in Section 1.

Experiments. We carried out two runs with 10 million iterations. It took about 2min30 for one run (60.000 iterations per second). The value of the estimator of the `icc` for the two runs is plotted as a function of the number of iterations in Figure 2b. We see that the estimator converges as expected to the value of the `icc` of TS represented by a straight horizontal line (and which was computed exactly in the previous section). We also plotted the estimated standard deviation as a function of the number of iterations. To obtain it, we did one billion iterations. We then estimated the standard deviation σ, and plotted $\frac{\sigma}{\sqrt{n}}$. We see that large jumps or discontinuity happen, but only at the beginning. They correspond to the draw of a fork with a lot of K22s and open K22s corresponding to a user who does not have the same `icc` as the global network. Then, the convergence is quick. After 300 iterations, the standard deviation is below 10% and after 1000 iterations, we do not experience a value of the runs less precise than 10%.

4 Results: Clustering coefficients in real Datasets

Twitter. To compute the number of K22s, open K22s, directed triangles, and undirected triangles in the Twitter Snapshot, we used a cluster with a rack of 16 Dell C6420 dual-Xeon 2.20GHz (20 cores), with 192 GB RAM, all sharing an NFS Linux partition over Infiniband. It took 51 hours to compute the exact numbers of K22s and open K22s, corresponding to 265h of cumulative computation times on the cluster. We reported the results in Table 1.

Number of K22s and triangles. We see that the numbers of K22s and open K22s are huge, 2.6×10^{16} and 3.1×10^{18}, respectively. It has to be compared with the number of triangles which are several orders of magnitude smaller: e.g., 2.5×10^{12} and 1.3×10^{14} for transitive triangles.

Clustering coefficient in the mutual graph. The mutual graph captures the friendship relationships in the social network. The mutual clustering coefficient thus is high ($mcc = 10.7\%$), as cliques of friends are frequent in Twitter.

Clustering coefficients in the whole graph. We observe that $\texttt{icc} = 3.3\% > tcc = 1.9\% > ccc = 1.7\% > ucc = 0.11\%$. Directed metrics better capture the interest relationships in the TS as ucc is very low. The highest parameter is the `icc`. It confirms the hypothesis of this paper that common interests between two users are better captured by the notion of K22 than by a direct link between these users. As expected, the second parameter is the one using transitive triangles. Indeed, they capture a natural way for a user of finding a new interesting user,

	icc	*tcc*	*ccc*	*ucc*
Twitter	3.1%	0.51%	0.24%	0.057%

Table 2: Clustering coefficients without the mutual structures.

that is, considering the followings of a following, especially after having seen retweets. A bit surprisingly, the ccc is not very low. In fact, a large fraction of the cyclic triangles are explained by corresponding triangles in the mutual graph (triangles of bi-directional links).
We believe bidirectional links contain a part of the social aspect of Twitter. Indeed, two friends will tend to follow each other, while a celebrity have little chance to follow back a person she does not know. A way to artificially take off the social influence in order to focus exclusively on the directed interest part of the graph is to remove the (open and closed) triangles and K22s contained in the mutual graph from the total count. Indeed, each undirected triangle of the mutual graph induces two cyclic triangles and six transitive triangles, and each undirected open triangle induces two open triangles. In the same way, each undirected K22 induces two K22s and each undirected open K22 induces two open K22s. The obtained results are shown in Table 2. If we take off those mutual triangles, both the tcc and the ccc values drop to 0.51% and 0.24%, respectively, while the `icc` stays about the same at 3.1%. This tends to confirm the hypothesis that the directed triangle clusterings somehow measure the friendship part of the TS more than the interest part. We also looked at the distributions of K22, open K22 and `icc` for each node, using definitions adapted from the triangle ones [19]. We obtain a value of 7.7% for the local `icc`. More details can be found in [16].

Other networks. We computed the different metrics on four other directed networks: two social networks, a web network and a citation network. The main dataset characteristics (more details in [16]) as well as their clustering coefficients are reported in Table 3. The main takeaways are the following:
- A high value of `icc` indicates the presence of clusters of interests such as research communities or interest fields.
- A high value of tcc is the sign of an important *local* phenomena of neighbors' recommendations and/or of a high hierarchical structure in the dataset.
- The ccc has no real social meaning. If its value can be high in a directed graph, this is only due to the presence of bidirectional arcs and triangles.
- Directed networks have a high *mcc*. Indeed, their bidirectional parts (mutual graph) have strong social communities, leading to a high clustering coefficient.
- The ucc is usually significantly lower, showing that the directed part of the network is better understood using directed clustering coefficients.
- Directed social networks have similar mixes of values of their undirected and directed clustering coefficients, however, with some notable differences, due to their diverse usages and information.

Additional work. As a complement of this study, we also proposed two applications using the K22s, which can be found in [16]:
- Recommendations. We propose to use the K22s to carry out link recommendation, as we advocate that the interest clustering coefficient is a good

	Is a SN	N	$\lvert E\rvert$	$\frac{\lvert E\rvert_m}{\lvert E\rvert}$	icc	tcc	ccc	mcc	ucc
Instagram	Yes	4.5×10^4	6.7×10^5	11%	12.0%	15.4%	3.7%	22.6%	4.1%
Flickr	Yes	2.3×10^6	3.3×10^7	62%	12.4%	12.2%	9.3%	13.9%	10.8%
Web (.edu)	No	6.9×10^5	7.6×10^6	25%	46.3%	59.6%	18.8%	78.5%	0.69%
Citations	No	3.8×10^6	1.7×10^7	0%	22.3%	9.1%	0%	*(none)*	6.7%

Table 3: Information and clustering coefficients of the directed datasets. SN means Social Network. N is the number of nodes, $|E|$ the number of edges, and $\frac{|E|_m}{|E|}$ the fraction of edges implied in a bidirectional link.

measure of common user interests. The principle is to recommend links closing a large number of K22s (instead, classically, of triangles). We discuss the strengths/weaknesses of this method for a set of Twitter users.

- Models with addition of K22s. We propose a new directed random growing model, based on the one introduced by Bollobas et al. [3], with addition of K22s. This way, the model is able to build random networks with a high `icc`, in order to represent the high values of `icc` found in real-world networks as presented in this paper. We prove that the in- and out- degree distributions of this model follow power-laws, as most of real-world networks, and show empirically the high value of the `icc`.

5 Conclusion, Discussion, and Future Work

In this paper, we introduce a new metric, the *interest clustering coefficient*, to capture the interest phenomena in a directed graph. Indeed, the classical undirected clustering coefficient apprehends the social phenomena that my friends tend to be connected. However, it is not adequate to take into account directed interest links. The interest clustering coefficient is based on the idea that, if two people are following a common neighbor, they have a higher chance to have other common neighbors, since they have at least one interest in common. We computed this new metric on a network known to be at the same time a social and information media, a snapshot of Twitter from 2012 with 505 million users and 23 billion links. The computation was made on the total graph, giving the exact value of the interest clustering coefficient, and using sampling methods. The value of the interest clustering coefficient of Twitter is around 3.3%, higher than (undirected and directed) clustering coefficients introduced in the literature and based on triangles, which we also computed on the snapshot.

Since both `icc` and classical `cc` represent a probability to close a structure, comparing their values with each other makes sense. However, an additional comparison with random models would be good to quantify what a "high value" means. In a network built with the $G(n, p)$ model, both directed `cc` and `icc` are equal to p. For the Twitter Snapshot, $p \sim 10^{-7}$, way smaller than the found values. An interesting study would be to compute those metrics on preferential attachment models; we keep this for a future work.

The `icc` is defined here for unweighted networks: it would be interesting to generalize it to weighted ones as a future work. We also would like to further investigate link recommendation based on the K22 structure defined for the interest clustering coefficient. Indeed, as there are several order more K22s than

triangles in social networks, it could lead to a more diverse recommendation system, than the ones based on triangles. It would be interesting to carry out a real-world user case study to investigate if users are more satisfied by such recommendations.

References

1. Noga Alon, Raphael Yuster, and Uri Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.
2. Alain Barrat and Martin Weigt. On the properties of small-world network models. *The Eur. Phys. Journal B-Condensed Matter and Complex Systems*, 13(3), 2000.
3. Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *ACM-SIAM symposium on Discrete algorithms*, 2003.
4. P Oscar Boykin and Vwani P Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, 2005.
5. Giorgio Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76, 2007.
6. Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying social networks at scale: macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 277–288. ACM, 2014.
7. Mark S Granovetter. The strength of weak ties. In *Social networks*. Elsevier, 1977.
8. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proc. of the 19th Intl. conference on World wide web*. ACM, 2010.
9. Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407(1-3):458–473, 2008.
10. Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *ACM IMC*, pages 29–42, 2007.
11. Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proc. of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.
12. Seyed-Vahid Sanei-Mehri, Ahmet Erdem Sariyuce, and Srikanta Tirthapura. Butterfly counting in bipartite networks. In *ACM SIGKDD*, pages 2150–2159, 2018.
13. Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E*, 75(2), 2007.
14. Hamza Shaban. https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/.
15. Nitai B Silva, Ren Tsang, George DC Cavalcanti, and Jyh Tsang. A graph-based friend recommendation system using genetic algorithm. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7. IEEE, 2010.
16. Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, and Stéphane Pérennes. Interest clustering coefficient: a new metric for directed networks like twitter. *arXiv preprint arXiv:2008.00517*, 2020.
17. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.
18. Jia Wang, Ada Wai-Chee Fu, and James Cheng. Rectangle counting in large bipartite graphs. In *2014 IEEE International Congress on Big Data*. IEEE, 2014.
19. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.