Analysis of geolocalized social networks based on simplicial complexes

Riccardo Fellegara
University of Maryland
College Park (MD), USA
felle@umd.edu

Ulderico Fugacci
University of Maryland
College Park (MD), USA
fugacci@umd.edu

Federico Iuricich
University of Maryland
College Park (MD), USA
iurif@umd.edu

Leila De Floriani
University of Maryland
College Park (MD), USA
deflo@umiacs.umd.edu

ABSTRACT

A common issue in network analysis consists in the detection and characterization of the key vertices and communities. To this purpose, visualization tools could be of great help to support domain experts in analyzing this kind of data. However, the size of real networks can seriously affect the practical usage of these tools, thus, requiring the definition of suitable simplification procedures that preserve the core network information. In this work, we focus on geolocalized social networks, and we describe a tool for the analysis of this kind of data based on topological information. Supported by recent trends in network analysis, our approach uses simplicial complexes as a model for social networks. A homology-preserving simplification is used for dealing with the data complexity and for reducing the information to be visualized to the essential. By combining the representation based on simplicial complexes and the simplification tool, we can efficiently retrieve topological information useful for the network analysis. Both the effectiveness and scalability of our approach are experimentally demonstrated.

Keywords

Network analysis and visualization; geolocalized social networks; homology-preserving simplification; simplicial complexes.

1. INTRODUCTION

Network analysis is an active research topic with a variety of applications including sociology, physics, electrical engineering, biology, and economics. A social network is a complex system consisting of individuals connected by specific relationships, such as friendship, common interests, and shared knowledge. The most common way to model a network is through a graph $G = (N,A)$ where each individual is represented as an element in the set of nodes N, while a friendship tie between two individuals is modeled through an arc of G. Different attributes are attached to vertices and edges depending on the type of the network. For example, when working with *location-based social networks* [18, 29], GPS coordinates are attached to each vertex of G, representing the geospatial position of the content that an actor decided to share (pictures, restaurant reviews, etc.). The level of complexity reached by this kind of representations is well known, and a vast literature exists dealing with the problem of extracting information using tools rooted in graph analysis [14].

In this work, we approach the problem of social network analysis from a different perspective. When using a graph G to encode the relations among the actors, we can only represent pairwise friendships. Here, we involve the use of simplicial complexes for representing more relations. Starting from the graph G, we will compute the cliques of G, defined as a set of actors with mutual friendship relations. Our objective here is twofold: (i) by using a suitable data structure for encoding the simplicial complex we reduce the number of entities stored (when the number of cliques is lower than the number of edges) and (ii) we extract structural information about the complex (such as the distribution of the cliques or the minimal non-faces) that would be impossible to compute on the graph. A minimal non-face, or blocker, roughly represents a small hole in the complex, corresponding to missing friendship relations. The simplicial complex computed on the graph G is initially free of blockers. By using an homology preserving simplification procedure, we are able to simplify the original simplicial complex reducing its size while preserving all the holes. Each one of these holes will be represented by a new blocker in the simplified representation.

We consider here specific networks in which users are provided with a geographical location. We denote such networks as *geolocalized social networks* since we are only considering a single location per user. Each edge of the original graph is also represented as "weighted" by computing the distance between the locations of the connected users. The same approach can be extended to a wider class of networks such as social, sensor, collaborative and biological networks.

The main contributions of this paper are:

- a set of topological queries for retrieving information from a network encoded as a simplicial complex;
- the description of the data structure used for encoding the simplicial complex and the implementation of these queries on it;
- the description of the simplification algorithm [16] used for creating blockers on a simplicial complex and the definition of an efficient algorithm for identifying them.

The remainder of this paper is organized as follows. In Section 2, we review some background notions on simplicial complexes

LBSN 16, October 31-November 03 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/3021304.3021309

and simplicial homology. In Section 3, we discuss related work on complex network analysis and on dimension-independent representations for simplicial complexes. In Section 4, we describe how we extract a simplicial complex from social network with the Stellar tree and how we encode the complex within it. In Section 5, we describe the network properties and degrees, and how to extract them from a simplicial complex using the Stellar tree. In Section 6, we describe our method for simplifying a geolocalized social network. In Section 7, we describe how to detect the blockers of a simplicial complex, introducing an algorithm for extracting them. In Section 8, we present experimental evaluations of our claims. Finally, concluding remarks are drawn in Section 9.

2. BACKGROUND NOTIONS

In this section, we introduce some background notions which are at the basis of our work, namely simplicial complexes and simplicial homology.

A *simplicial complex* Σ on a finite set V is a collection of non-empty subsets of V, called *simplices*, such that if $\tau \in \Sigma$, $\sigma \subseteq \tau$, then $\sigma \in \Sigma$. Given a simplicial complex Σ, the elements of V are called *vertices* of Σ and a simplex $\sigma \in \Sigma$ is called a *k-simplex* if it consists of $k+1$ vertices. In the following, we denote a k-simplex $\sigma = \{v_0, v_1, \ldots, v_k\}$ as $v_0 v_1 \ldots v_k$ and each non-empty subset of σ as a *face* of σ. Geometrically, each simplicial complex can be considered as a subspace of a suitable Euclidean space $\mathbb{E}^n$ and, in such a context, each k-simplex is represented as the convex hull of $k+1$ geometrically independent points. For instance, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron and so on. Given a k-simplex σ, the *dimension of* σ is defined to be k and denoted as $dim(\sigma)$. More generally, in the following we will use the term dimension to denote the size decreased by one of any subset of V even if it does not represent a simplex of Σ. The *dimension of a simplicial complex* Σ, denoted as $dim(\Sigma)$, is the largest dimension of the simplices in Σ. For example, Figure 1(b) shows a simplicial 3-complex since its biggest simplex is a tetrahedron, i.e. σ_2. For $k \geq 0$, the *k-skeleton* of Σ is the simplicial complex $\Sigma^{(k)}$ consisting of the simplices of Σ of dimension less or equal to k. Given a simplex $\sigma \in \Sigma$, the *star* of σ, denoted as $St(\sigma)$, is the set of the simplices of Σ containing σ. A simplex $\sigma \in \Sigma$, for which $St(\sigma)$ consists only of σ itself, is called *top simplex* (or, equivalently, *facet*). In the following, we denote as $St_{top}(\sigma)$ the set of the top simplices of Σ containing σ. For the simplicial complex shown in Figure 1(b) the set $St_{top}(v) = \{\sigma_1, \sigma_2, \sigma_3\}$. The *link* of a simplex $\sigma \in \Sigma$, denoted as $Lk(\sigma)$, is the simplicial complex consisting of the faces, having empty intersection with σ, of the simplices in $St(\sigma)$.

In this work, we will consider simplicial complexes obtained by expanding the cliques of a graph, also called *flag complexes*. A *flag complex* of a graph $G = (V, E)$ is the simplicial complex $Flag(G)$ whose simplices coincides with the cliques of G. In this framework, the top simplices of $Flag(G)$ correspond to the maximal cliques of G and the 1-skeleton of $Flag(G)$ is the graph G itself. The simplicial complex depicted in Figure 1(b) is the flag complex computed on the graph G depicted in Figure 1(a).

Studying the topology of a flag complex is important for retrieving its structural properties. Some of these properties can be obtained by retrieving the holes (or, more formally, the homology) and the blockers of the associated simplicial complex.

Simplicial homology is a powerful tool in shape analysis, providing topological invariants for shape description. Roughly speaking, homology reveals the presence of "holes" in a shape. In dimension 0, these represent the connected components of the complex, in dimension 1, its tunnels and its holes, in dimension 2, the shells surrounding voids or cavities, and so on.

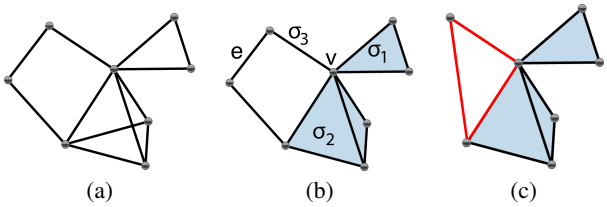

Figure 1: (a) A graph and (b) the corresponding flag complex. After removing edge e from the flag complex, a blocker (in red) is created (c).

The notion of blocker, introduced in [1], is related to the notion of hole in the considered simplicial complex. Given a simplicial complex Σ on V, a *blocker* (or, equivalently, a *minimal non-face*) of Σ is a subset σ of V of dimension strictly greater than 1 such that every subset of σ except for σ itself is a simplex of the simplicial complex Σ. In Figure 1(c), the only blocker of the simplicial complex is the missing triangle having all its three edges (depicted in red) part of Σ.

3. RELATED WORK

In this section, we provide an overview of the methods, based on simplicial complexes, for analyzing networks. Then, we present the state-of-the-art data structures commonly used for encoding simplicial complexes of arbitrary dimension.

3.1 Complex network analysis

Complex network analysis concerns the study of systems representing connections between distinct elements or actors [24, 28, 30]. Networks have become a useful tool to represent systems from a wide variety of research fields. Examples include social, sensor, collaborative and biological networks [7].

Several methods have been proposed to analyze a network from different viewpoints. The term *egocentric* networks, refers to the analysis of a network focusing on the study of the ties of a single individual and on its local relevance. Thanks to suitable *centrality measures*, this kind of analysis can identify different roles for an individual such as *isolated*, *outlier*, *broker* or *keyplayer* [8].

Sociocentric (or, *whole*) networks focus on large groups of individuals or elements studying global and structural properties of the entire network. The connectivity of a network can be measured through a large variety of attributes and descriptors such as density, cohesion, diameter, small worlds, bridges and structural holes. A relevant issue in this kind of analysis is related to the study of *communities* such as groups or social circles [13, 21]. In this respect, a core notion is captured by the concept of *clique* [20].

Based on these notions, different methods, which exploit statistical, combinatorial or topological techniques, have been developed allowing to retrieve the connectivity structure of a network. The retrieved information is usually collected and visualized thanks to *clustered* or *dynamical representations*. In the first case, the individuals of a network are partitioned with respect to the communities they belong to; in the other one, portions of the network are dynamically highlighted according to evolving parameters reflecting network cohesion.

Analyzing networks through simplicial complexes

Different proposals can be found using simplicial complexes for analyzing a network. Including all the information provided by the original graph, the simplicial complex has been used for computing

descriptors that better reflect relations among data, representing communities of strongly-connected individuals.

Applications involve networks of various nature including collaborative [22, 31, 4], social and communication [17], sensor [9, 23, 6], multi-radio multi-channel [26] and random [15] networks.

In these contexts, the dimension k of the simplices has been studied for characterizing the pairwise connections among individuals. Moreover, the centrality of an individual and the cohesion of a network can be expressed in terms of adjacency and incidence relations between simplices. Other advantages in the use of simplicial complexes are the possibility of detecting and localizing connectivity holes in a network, and that the notions of homology and blocker are available.

3.2 Data structures for simplicial complexes

Working with flag complexes, no assumptions can be made on the dimension of the simplicial complex obtained. For this reason, the data structure adopted must be dimension-independent. Most topological data structures for simplicial complexes can only represent complexes in a specific dimension (generally two or three).

We can identify four dimension-independent data structures commonly used in the literature for representing a simplicial complex.

Given a simplicial complex Σ, the *IG* [11] is a data structure describing its Hasse diagram [25], i.e., the graphical representation of the partially ordered set generated by all the simplices of Σ and their incidence relations. A graph is used for associating the simplices of a simplicial complex Σ with the nodes of the graph itself while the boundary and coboundary relations between the simplices of Σ are associated with its arcs.

As the *IG*, the Simplex Tree (*ST*) [2] encodes all the simplices of Σ and it has been defined with the purpose of limiting the number of incidence relations encoded. By using the Simplex Tree the storage consumption required by encoding the entire graph has been reduced, though encoding one node for each simplex in Σ does not guarantee scalability to higher dimensions. To this end, data structures based on the encoding of only top simplices and vertices have been shown to be particularly effective.

The *Generalized Indexed data structure with Adjacencies (IA*)* [3], for example, has been shown to be exceptionally compact encoding only the vertices and top simplices of a simplicial complex Σ, plus a subset of its adjacent and boundary relations. For each top k-simplex τ, the relation with its vertices is encoded as well as the relations to all the top simplices sharing a face of dimension $k-1$ with τ. The *Stellar tree* [12] enhances this idea encoding exclusively the relation of a top simplex with its boundary vertices and providing a mechanism to efficiently extract all the other topological relations at runtime. We refer to the next section for further details.

4. REPRESENTING A NETWORK ON THE STELLAR TREE

The practical relevance of our work is strongly connected to the data structure we have chosen for representing the simplicial complex, the Stellar tree. A *Stellar tree* [12] is a spatio-topological data structure that uses a hierarchical structure for representing the embedding space of a simplicial complex. Given a simplicial complex Σ, a point-region quadtree [27] is created based on the vertices of Σ. A single parameter, denoted as k_v, determines the maximum number of points indexed by a leaf of the quadtree. Each leaf of the quadtree, also called *leaf block*, encodes the indexes of the contained vertices alongside with the indexes of the top simplices incident in those vertices. This formulation enables the usage of a very simple (indexed) representation for Σ and defers the decision of the topological data structure, and its encoded connectivity relations, to runtime.

From a graph to a simplicial complex

When developing a suitable representation for flag complexes computed on social networks, our input complex is a graph describing the friendship relations between vertices. Starting from this graph, the top simplices are obtained computing the maximal cliques. The Stellar decomposition can be used to this purpose for enhancing performances [16].

The procedure iteratively visits the leaf blocks and applies on the local graph contained in each leaf the Bron-Kerbosch algorithm with pivoting [5]. We identify as *local graph* the set of edges having at least one vertex contained in the leaf block currently visited and the relative vertices (both inside and outside the leaf). As described in [5], following this approach the maximal cliques formed by vertices belonging to adjacent leaf blocks are identified multiple times, namely one for each leaf block. However, a clique is inserted into the top simplex array only when we are processing the leaf block indexing its vertex with maximum index.

5. COMPUTING CENTRALITY INFORMATION FROM THE SIMPLICIAL MODEL

In this section, we define the information we are going to extract from a flag complex Σ. The most basic value that we are going to consider is the *vertex degree*, equivalently defined as the number of connected edges or the number of adjacent vertices. This notion is extended in [22] to the analysis of co-authorship networks defining the degree of a k-simplex. In this context, the degree of a vertex represents the number of distinct co-authors that collaborated with him/her.

A simplex represents a set of authors having at least one joint work together. The simplex degree represents the number of distinct co-authors who have jointly published with them. In our case, the relevance of the simplex degree is maintained but its meaning is different since our network is based on friendship relations. For each k-simplex σ we have:

- *lower* degree: number of $(k-1)$-simplices contained in σ (constantly equal to $k+1$);
- *upper* degree: number of $(k+1)$-simplices containing σ;
- *top* degree (also called *facet* degree): number of top simplices containing σ.

While the upper and lower degrees define a measure for evaluating the relations between cliques, the facet degree specifically considers relations with the maximal cliques (facets) only. Thus, given a cluster of friends σ, its *lower degree* expresses the importance of the clique in terms of number of individuals in the community. Dually, the *upper degree* of σ underlying the importance of the community σ as the number of individuals V_σ having a friendship relation with all the vertices in σ. Considering the vertices in V_σ, no information on the mutual friendships is provided by the upper degree. However, this can be retrieved by focusing on the maximal cliques. The *top degree* of a simplex σ is defined as the cardinality of the top simplices in the star of σ (i.e., $|St_{top}(\sigma)|$), and thus, corresponds to the number of communities to which all the vertices of σ belong. The higher the top degree is, the more σ is connected with other communities in the complex. Since we are retrieving facets, for each community we already know the connectivity of its vertices and we can compute how much strong such community is (lower degree).

5.1 Computing network properties on the Stellar tree

Once we compute the top simplices on the Stellar tree, as described in Section 4, we can extract any other topological relation. The basic paradigm for using the Stellar tree is to locally process the simplicial complex in a streaming manner by iterating through the leaf blocks. For each leaf block b, we extract a local application-dependent data structure, that is discarded once we have processed b. This allows for allocating the memory, required for representing the topological structures, for a single leaf block at a time. Moreover, queries are processed locally and the cost of computing the topological relations is amortized over multiple accesses while processing each leaf block.

The descriptors that we want to explicitly encode are the top and upper degrees for each vertex. The upper degree will be useful to evaluate the vertices that have the higher number of friends while the top degree is used to identify vertices connected to the higher number of communities. As we explicitly represent the vertices and the top simplices of Σ, all the degrees can be computed extracting the top simplices in the star of each vertex. Thus, we have practical advantages for computing these degrees by using the Stellar tree.

Within each leaf block b, the algorithm iterates through the vertices of the top simplices in b. For each contained vertex v, we encode the top simplices in $St_{top}(v)$ and the edges in $St(v)$. Then, once we have extracted these local topological relations, we compute the local degrees, and we update accordingly the global ones. The upper degree of each vertex v is equal to the number of edges in its star while the top degree of v is the size of its $St_{top}(v)$. The complexity of the algorithm within each leaf block b is linear to the number of the top simplices in b. Finally, these two integers values are stored in a global structure and the local topological relations are discarded before processing the next leaf blocks.

6. SIMPLIFYING THE SIMPLICIAL MODEL

Simplifying the structure of a social network is a valid approach for disclosing new insight on the data. Since, we are interested in modeling the network as a simplicial complex, we need to preserve the structural information that such complex can provide. A way for preserving these information is to adopt an homology-preserving simplification procedure.

Edge contraction is the most common operator for simplifying simplicial complexes. It has been used in computer graphics and visualization and more recently in topological data analysis for reducing the size of higher dimensional simplicial complexes [1]. Given a simplicial complex Σ, let us consider a pair v_1,v_2 of its vertices forming an edge $e = v_1v_2$ of Σ. The edge contraction consists of the collapse of one vertex of e on the other one. Without loss of generality, in the following we denote v_1 as the vertex collapsing on v_2 and, consequently, v_2 as the surviving vertex. It can be formally described as a function mapping a simplex σ of Σ in the simplex $\mu(\sigma)$ spanned by the vertices of a σ in which vertex v_1 has been replaced by vertex v_2. As a result of the edge contraction, all the simplices in $St(e)$ are removed from Σ and each simplex σ in $St(v_1) \setminus St(e)$ is redirected into $St(v_2)$ by mapping σ into $\mu(\sigma)$. Edge contraction is an operation linear in the number of simplices in $St(v_1) \cup St(v_2)$.

One of the advantages in using the Stellar tree is that of exploiting the representation based on the top simplices to outperform normal simplification algorithms. Using the algorithms defined in [16], we can perform the edge contraction only working on the top simplices in $St_{top}(e)$, i.e., the set of top simplices incident in the edge removed, and in $St_{top}(v_1)$, i.e., the set of top simplices incident in the vertex removed with e. The above algorithm successfully reduces the size of a simplicial complex Σ but, in general, does not preserve the homology of Σ. The *link condition*, introduced in [10] for a 2- or 3-manifold and extended in [1] to arbitrary simplicial complexes, defines a constraint for the edge contraction applied to a simplicial complex Σ for preserving its homology. Given a simplicial complex Σ, an edge $e = v_1v_2$ of Σ satisfies the link condition if and only if $Lk(v_1) \cap Lk(v_2) = Lk(e)$. Again here, we can use the top simplices for efficiently checking this condition following the algorithm described in [16].

Based on homology-preserving edge contractions, we develop a simplification algorithm for geolocalized social networks. Given a geolocalized social network modeled by a graph $G = (V,E)$, let us considered the flag complex $\Sigma = Flag(G)$ of G. The simplification algorithm consists in performing homology-preserving edge contractions whenever there are edges satisfying the link condition. In order to maintain the topological structure of a network and, at the same time, to extract its core information, the simplification process can be driven by semantic criteria.

Given an edge $e = v_1v_2$ of Σ (or, equivalently, of G), we can label it according with two different weights. Thanks to the GPS coordinates associated with each vertex of Σ, e can be weighted by the *spatial distance* between the geo-spatial positions of its vertices. Alternatively, we can define a weight based on the list of check-in associated with each vertex of Σ. For $i = 1,2$, we denote as L_i the set of the locations at which v_i did a check-in. We define the *check-in weight* between v_1 and v_2 as

$$\frac{|L_1 \setminus L_2| + |L_2 \setminus L_1|}{|L_1 \cup L_2|}$$

This weight assumes values between 0 and 1. Values close to 0 occurs when the two check-in lists are very similar, while values close to 1 denotes that the two lists are almost disjoint.

Once we have chosen a criterion for assigning a weight to the edges, the latter are processed in ascending order and contracted if they satisfy the link condition. Given an edge $e = v_1v_2$ to be contracted, we choose as surviving vertex v_2 the one with higher vertex degree in the original complex Σ.

7. EXTRACTING BLOCKERS FROM THE SIMPLICIAL MODEL

As underlined in the literature, the notions of homology and blocker have been pointed out as relevant tools for detecting missed connections between individuals who could potentially set up a community. Thus, the detection of the blockers of a simplicial complex Σ is a key task for the analysis of the network.

Equivalently to the definition in Section 2, a blocker of Σ is a minimal (w.r.t. inclusion) simplex in the set difference $Flag(\Sigma^{(1)}) \setminus \Sigma$. This ensures that each blocker of Σ is necessarily contained in a maximal clique of $\Sigma^{(1)}$ which does not belong to Σ. In order to reveal the blockers contained in such a maximal clique σ of $\Sigma^{(1)}$, the procedure verifies if all the subsets of dimension $k-1$ of σ belong to Σ, otherwise it recursively verifies for each subset τ of σ of dimension $k-1$ which does not belong to Σ. By retrieving all the maximal cliques σ of $\Sigma^{(1)}$ not belonging to Σ and by executing the procedure on each of such σ, it is possible to detect all the blocker of Σ.

Taking in input the maximal clique $\sigma = v_1v_2v_3v_4$ of the simplicial complex Σ depicted in Figure 2, the procedure considers all its subsets of dimension two. Triangles $v_1v_2v_3$, $v_1v_3v_4$ do not belong to Σ, so the procedure is executed on each of them returning that

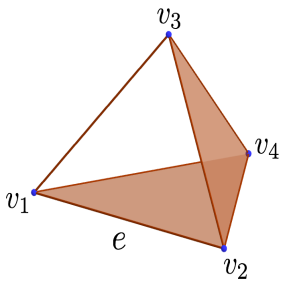

Figure 2: Some relevant elements of a simplicial complex Σ: tetrahedron $v_1v_2v_3v_4$ is a maximal clique; triangle $v_1v_2v_3$ is a blocker which coincides with a missing simplex of edge $e = v_1v_2$.

the two triangles are blockers of Σ.

In real cases, the latter procedure will often analyze maximal cliques of high dimensions even if the actual blockers have very low dimensions. Since in the worst case we have to consider all possible subsets of σ, whose number grows exponentially when the dimension of σ increases, the fact that the maximal cliques dimensions often assume high values leads to expensive computations.

To overcome this limitation and obtain the blockers of Σ more efficiently, we execute the algorithm on a collection of subsets of V of lower dimensions with respect to the ones above considered. Given a simplicial complex Σ and an edge $e = v_1v_2$ of Σ, we denote T_1 (and similarly T_2) as $St_{top}(v_1) \setminus St_{top}(e)$, i.e., the set of top simplices which are in the star of v_1 but not in that of e. We say that e admits a missing simplex if there exist τ_1 in T_1 and τ_2 in T_2 such that $\rho := \tau_1 \cap \tau_2 \neq \emptyset$ and $St_{top}(\rho) \cap St_{top}(e) = \emptyset$. Under such assumptions, we refer to $\rho \cup e$ as a *missing simplex of* e. By definition, each missing simplex of an edge of Σ does not belong to Σ.

By considering the edge $e = v_1v_2$ in Figure 2, we have that the set of top simplices T_1 and T_2 consist of edge v_1v_3 and of triangle $v_2v_3v_4$, respectively. So, since the set of top simplices which are in the star of both e and v_3 is empty, the empty triangle $v_1v_2v_3$ is a missing simplex of e. The following proposition ensures that considering the set of missing simplices of the edges of Σ guarantees the retrieval of the set of blockers of Σ.

PROPOSITION 7.1. *Let Σ be a simplicial complex. Then, for each blocker σ of Σ, there exists a missing simplex containing σ.*

PROOF. Given any edge $e = v_1v_2$ contained in σ, let us consider the blocker σ as the disjoint union $\sigma = \tau \cup e$. Since σ is a blocker, we have that, for $i = 1,2$, the simplex $\tau \cup v_i$ belongs to Σ. Let τ_i be a top simplex of Σ containing $\tau \cup v_i$. Simplex τ_i does not contain e otherwise σ belongs to Σ. By considering $\rho := \tau_1 \cap \tau_2$, we have that ρ is non-empty since it contains τ. Further, $St_{top}(\rho) \cap St_{top}(e) = \emptyset$. In fact, the existence of a top simplex τ' of Σ containing both ρ and e implies that τ' is in the star of $\sigma = \tau \cup e$ leading to a contradiction. As a consequence, one can claim that $\rho \cup e$ is a missing simplex of e containing σ. □

Similarly to the maximal cliques of $\Sigma^{(1)}$ not belonging to Σ, the missing simplices of the edges of Σ are in $Flag(\Sigma^{(1)}) \setminus \Sigma$. This ensures that each missing simplex contains at least one blocker. Differently to the maximal cliques, the missing simplices are not, in general, maximal elements with respect to the inclusion of that set difference. As we will discuss in Subsection 8.2, the relative low dimensions of the missing simplices allow to improve the computational times of the blocker extraction.

Table 1: The table shows the simplification statistics, as the number of vertices $|V|$, the number of top simplices $|T|$ and the dimension d in the initial and in the final simplicial complex, the simplification ratio (%), plus the experimental memory peaks (expressed in megabytes) obtained by the *Star* data structure (*star*) and by the Stellar tree (*tree*).

	weight	d.s.	$\|V\|$	$\|T\|$	d	%	mem.
OCEANIA	input complex		1.99K	4.62K	13		
	geo-dist.	star	0.32K	1.78K	10	84	8.20
		tree	0.35K	1.87K	11	82	6.73
	check-in	star	0.33K	2.32K	11	83	8.27
		tree	0.36K	1.77K	11	82	6.48
EUROPE	input complex		9.88K	18.4K	12		
	geo-dist.	star	1.25K	7.55K	8	87	23.4
		tree	1.42K	7.35K	10	86	16.7
	check-in	star	1.17K	8.37K	10	88	24.0
		tree	1.43K	7.56K	9	86	16.6
N. AMERICA	input complex		31.4K	90.3K	18		
	geo-dist.	star	2.78K	52.6K	18	91	97.1
		tree	3.37K	47.4K	16	89	56.4
	check-in	star	2.64K	55.7K	18	92	102
		tree	3.37K	48.0K	17	89	60.8

8. EXPERIMENTAL RESULTS

In this section, we evaluate the performances of our approach on real geolocalized social networks. We have chosen the BRIGHTKITE network from the *SNAP Datasets Collection* repository [19]. Given this dataset, we have selected three sub-networks, localized in Europe, Norther-America and Oceania continents. On top of these networks, we generate the three corresponding flag complexes using the algorithm described in Section 4. These datasets contains from 2 to 30 thousands points, leading to simplicial complexes from 12 to 17 dimensions. Further, as the clique-based blocker computation procedure cannot be completed on the above simplicial complexes, we use also a smaller dataset based on the Souther-America continent, that has 6 hundreds points, leading to 9 hundreds top simplices in at most 6 dimensions.

The Stellar tree implementation is compared against a global structure, called in the following *Star* data structure. The Star data structure explicitly encodes, for each top simplex, the boundary relation to its vertices and, for each vertex v, the top simplices in the star of v. These are the two minimal connectivity relations for efficiently execute the simplification procedure. The hardware configuration used for these experiments is an Intel i7 3930K CPU at 3.20Ghz with 64 GB of RAM. The CPU can handle up to 12 threads in parallel.

8.1 Homology-preserving simplification

In this subsection, we evaluate the statistics of the Stellar tree and the global structure while simplifying a simplicial complex. Table 1 shows the compression factors and the memory peaks.

The simplification sequence that we obtain from the Stellar tree and on the *Star* data structure is generally different. The reason is that the Stellar tree uses a local priority queue for organizing the simplifications inside each leaf block while the global data structure only uses a global queue. We can notice from our results that the simplification order influences the number of contractions executed, and the Stellar tree always execute slightly fewer simplifications than the global data structure. The simplification ratio is, on average, around 80-90% of the initial vertices. Conversely,

Table 2: Comparison between clique-based and missing-based approaches. Table shows, for each dataset, the number of cliques, missing simplices and blockers (columns *num*), plus their maximum dimension (columns d). Columns *time* indicate timings of blocker extraction.

	weight	cliques-based			missing-based			blockers	
		num	d	time	num	d	time	num	d
SA	geo-dist.	0.34K	8	6.07s	2.72K	6	0.05s	1.06K	4
	check-in	0.32K	8	1.25s	2.18K	6	0.04s	0.94K	3
OC	geo-dist.	1.80K	12	>12h	24.9K	10	21.5s	6.10K	4
	check-in	2.14K	17	>12h	90.4K	12	1.14h	7.10K	3
EU	geo-dist.	11.0K	13	>12h	82.9K	9	13.8s	29.2K	4
	check-in	14.0K	13	>12h	133K	11	3.01m	31.6K	3

the dimension of the original complex is not significantly reduced. Moreover, the simplification ratio is rather similar also considering a distance-based or a check-in-based simplification, meaning that the homology preservation constraint is not significantly influenced by the weights associated with the network edges.

Finally, analyzing the memory peak, we can observe that the Stellar tree is always more compact than the *Star* data structure, requiring from 60% to 80% of the memory.

8.2 Blocker computation

In this subsection, we consider the blockers extraction evaluating the performances of the two strategies described in Section 7. We call *clique-based* strategy the approach that considers the set of the maximal cliques of $\Sigma^{(1)}$ not belonging to Σ. Conversely, we call *missing-based* the approach considering the missing simplices of the edges of Σ.

Table 2 shows the experimental results obtained indicating, for both maximal cliques and blockers, their number and their maximum dimension. Since the flag complex associated with a network is free of blockers, experiments involve datasets obtained after geo-distance or check-in based simplifications. We can notice that the number of missing simplices is always relevantly higher than the number of maximal cliques (from 7 to 42 times higher). Though, the missing-based approach is always extremely faster than the clique-based, as in most of the cases, the latter is killed after 12 hours of computation.

The blockers extraction slows down exponentially with the increasing of the dimension of the simplices given as input, and we observe that the maximum dimension of the candidates found with the missing-based approach is generally lower than the dimension of the candidates found with the clique-based approach. Further evidence of this trend can be found by comparing the distributions of the maximal cliques and of the missing simplices of the complex.

As depicted in Figure 3, the experimental evaluations highlight that, in practical cases, the distribution of maximal cliques is nearly uniform across the dimensions, while the distribution of the missing simplices, and the distribution of the blockers, are concentrated mainly in low dimensions. As discussed in Section 7, the procedure execution is time-consuming on simplices of high dimension, while on simplices of low dimension is pretty fast (constant-time for simplices of dimension 2). For this reason, in spite of the higher number of simplices to be analyzed, the concentration of the missing simplices in low dimensions ensures that the missing-based strategy is computationally efficient compared to the clique-based approach.

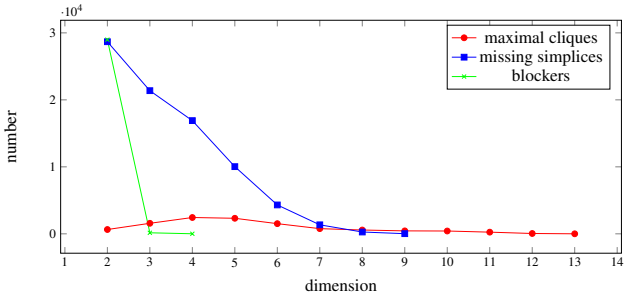

Figure 3: Comparison between the distributions of the maximal cliques, the missing simplices and the blockers of the simplicial complex obtained after the geo-distance based simplification of the EUROPE dataset.

Table 3: The table shows the network degrees of the input complex and of the simplified ones, following the geo-distance or the check-in based, with a *Star* data structure (*star*) or the Stellar tree (*tree*).

	weight	d.s.	1-skeleton		degrees			
					vertex		top	
			vertices	edges	avg	max	avg	max
OCEANIA	input complex		1.99K	6.12K	6.16	0.20K	4.79	0.99K
	geo-dist.	star	0.32K	2.32K	14.6	0.13K	3.83	0.46K
		tree	0.35K	2.21K	12.6	0.11K	4.20	0.46K
	check-in	star	0.33K	2.63K	15.9	0.16K	5.01	0.78K
		tree	0.36K	2.16K	12.0	0.13K	4.04	0.44K
EUROPE	input complex		9.88K	26.5K	5.37	0.24K	3.50	0.97K
	geo-dist.	star	1.25K	9.75K	15.6	0.24K	3.30	0.55K
		tree	1.42K	9.23K	13.0	0.21K	3.16	0.50K
	check-in	star	1.17K	10.9K	18.5	0.19K	3.76	0.69K
		tree	1.43K	9.34K	13.1	0.26K	3.24	0.59K
N. AMERICA	input complex		31.4K	112K	7.17	0.85K	4.34	10.4K
	geo-dist.	star	2.78K	60.5K	43.5	1.00K	4.11	6.39K
		tree	3.37K	55.3K	32.8	0.97K	3.28	3.73K
	check-in	star	2.64K	63.6K	48.1	0.78K	4.54	7.92K
		tree	3.37K	55.5K	32.9	0.94K	3.41	4.49K

8.3 Analyzing the network

In this subsection, we report the values obtained by computing the degrees, introduced in Subsection 5. The statistics are reported in Table 3 indicating the variation of the 1-skeleton between the initial and the simplified complexes, and showing how the vertex and top degrees are affected by the homology-preserving simplification process.

We can observe that the 1-skeleton is highly reduced by the simplification process, and the reduction ratio of the edges is proportional to the one of the vertices of the complex. Comparing the vertex degrees of the initial and simplified complex, we notice that the average of the simplified complex at least triples the initial average, meaning that the remaining vertices communities are strongly connected with the other communities in their neighborhood. This trend can be observed also considering the maximal values that, with the exception of OCEANIA dataset, remains high.

Considering the variation of the top degrees, we can note, from the value 0 of the minimum degree in the final complex (not shown in the Table), that all the datasets have more than one connected component. Notice that, if a connected component has no homology cycles it is correctly reduced to a single point by the simplification.

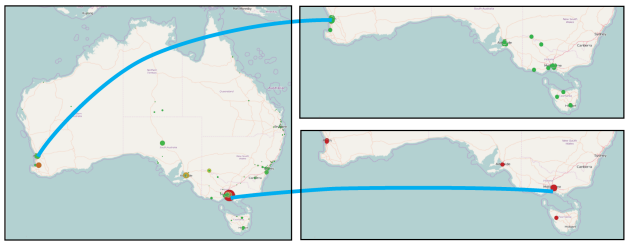

Figure 4: For the OCEANIA dataset (focus on Australia), vertices are depicted with increasing dimension according to the number incident top simplices and colored (from green to red) depending on the number of adjacent vertices.

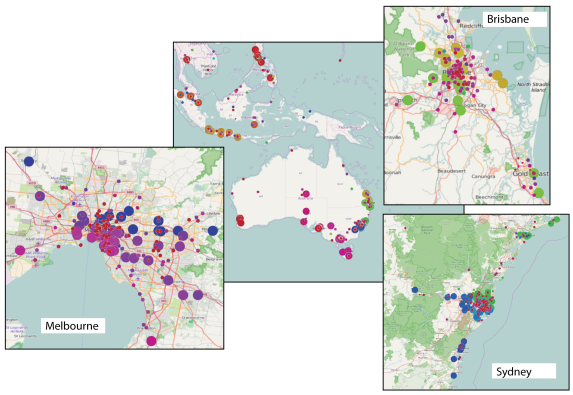

Figure 5: For the OCEANIA dataset (with a focus on three cities), the effects of the simplification process are shown depicting the correlation between vertices that survived the simplification process from those that have been contracted during it.

Visualization-aided network analysis

The information extracted from the simplicial complex are particularly useful when coupled with visualization tools. In Figure 4, we are focusing on *Australia* showing a dot for each vertex (edges are avoided for clarity). Dots are represented with increasing size, depending on the number of top simplices incident in each vertex, and colored (from green to red) depending on the number of adjacent vertices (upper degree).

The first information let us discriminate the different clusters grouped in the main cities. While the color underlines the amount of existent connections of a vertex, its size represents how many of these adjacent vertices are connected with each other and if they represent a strongly connected community. Looking at Figure 4 (left), for example, we can notice that in Melbourne we have a strong community mainly centered in a big vertex, in Perth the connectivity of the vertices is lower but we still have some connected communities while in Brisbane and Sydney the number of top simplices is lower (and thus the strength of the community is lower).

Going a step further, we also want to analyze how the simplification process influences the structure of the network. Using the geo-distance based simplification we notice two opposite changes in the network connectivity depending on the properties of the vertices. If the vertex was originally strongly connected and belonging to big clusters (i.e., the big red dot in Melbourne), then its neighborhood will be strongly simplified and the number of adjacent vertices will reduce drastically. An example can be seen in Figure 4 (lower right). Here, we are depicting only the vertex chosen for the analysis (big red vertex) and those connected to it at the end of the simplification process (small red dots). The vertex was connected to 201 vertices and 966 top simplices at the beginning of the process but now it is connected to only 5 vertices.

Dually, we consider the case where the vertex is loosely connected (for example, the green vertex in the higher right figure). At the beginning of the simplification process, it was connected to only 12 vertices, but after the simplification, it results connected to 96 different vertices.

This opposite behavior is a natural consequence of using the blockers to inhibit simplifications. When we are processing strongly connected communities, the amount of top simplices is also high and it is not possible to generate blockers. Conversely, processing vertices that are decently connected but not in the center of the network will generate much more blockers. The results obtained can actually be used to avoid that the big communities incorporate the smaller ones along the simplification process. Using the former approach we are able to identify the relations depicted in Figure 5.

For the entire dataset (and for the zoomed images of the three Australian cities), we show with bigger dots the vertices surviving the simplification process. These are then colored according to the area they belong to on the map. Dually, smaller dots represent the vertices contracted (and removed) inside one bigger dot. These are marked using the color of the vertex that survives the contraction. We can clearly see that even if some vertices are simplified with vertices belonging to the same region/city, there are actually many vertices that have been contracted with those far away from their original city.

9. CONCLUDING REMARKS

In this paper, we have presented a new approach for analyzing geolocalized social networks based on simplicial complexes. Starting from a graph representation of the original network, a simplicial complex is obtained as a collection of all the maximal cliques extracted from the graph.

The latter representation incorporates all the information provided by the original graph in adjunct to a qualitative description of the structural properties of the network. By means of the simplicial complex, we have been able to represent relations between the vertices of the network and their communities. Moreover, using an homology-preserving simplification process, we have been able to identify missing friendships that were not represented in the original graph. This has been done by studying the blockers, missing simplices representing the homology cycles of the original complex.

We are developing a fully interactive visualization tool coupling the information extracted from the original graph and those extracted from the simplicial complex. Data extracted from the two representations will provide complementary information for leading a user to the salient parts of the network. The final aim of this analysis is to incorporate this process in a friendship-suggestion procedure for empowering smaller communities of the network.

To achieve such a result, a fundamental step will be encoding the sequence of simplifications in a multi-resolution model. Based on the latter, the user will gain control on the resolution of the simplicial complex and will be able to study the evolution of the network interactively.

We are also considering the extension of this framework to the analysis of time-varying social networks. In this case, the proposed degrees are still computable but more involved tools could be preferable for studying the evolution of data over time. Specifically, we

are considering the use of the persistent homology as a replacement for the blockers.

Acknowledgments

This work has been supported by the US National Science Foundation under grant number IIS-1116747. The BRIGHTKITE dataset is courtesy of the *SNAP Datasets Collection* [19].

10. REFERENCES

[1] D. Attali, A. Lieutier, and D. Salinas. Efficient data structure for representing and simplifying simplicial complexes in high dimensions. *International Journal of Computational Geometry & Applications*, 22(04):279–303, 2012.

[2] J.-D. Boissonnat and C. Maria. The simplex tree: an efficient data structure for general simplicial complexes. *Algorithmica*, 70(3):406–427, 2014.

[3] D. Canino, L. De Floriani, and K. Weiss. IA*: an adjacency-based representation for non-manifold simplicial shapes in arbitrary dimensions. *Computers & Graphics*, 35(3):747–753, 2011.

[4] C. Carstens and K. Horadam. Persistent homology of collaboration networks. *Mathematical Problems in Engineering*, 2013, 2013.

[5] J. Cheng, L. Zhu, Y. Ke, and S. Chu. Fast algorithms for maximal clique enumeration with limited memory. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 1240, 2012.

[6] H. Chintakunta and H. Krim. Divide and conquer: localizing coverage holes in sensor networks. In *7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pages 1–8. IEEE, 2010.

[7] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[8] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: a survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[9] V. De Silva and R. Ghrist. Homological sensor networks. *Notices of the American mathematical society*, 54(1), 2007.

[10] T. K. Dey, H. Edelsbrunner, S. Guha, and D. V. Nekhayev. Topology preserving edge contraction. *Publ. Inst. Math.(Beograd)(NS)*, 66(80):23–45, 1999.

[11] H. Edelsbrunner. *Algorithms in combinatorial geometry*. Springer Verlag, Berlin, 1987.

[12] R. Fellegara. *A spatio-topological approach to the representation of simplicial complexes and beyond*. PhD thesis, Department of Computer Science (DIBRIS), University of Genova, 2015. Internal Report DIBRIS-TH-2015-01.

[13] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[14] O. Frank. A survey of statistical methods for graph analysis. *Sociological methodology*, 12:110–155, 1981.

[15] D. Horak, S. Maletić, and M. Rajković. Persistent homology of complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03034, 2009.

[16] F. Iuricich, R. Fellegara, and L. De Floriani. Efficient representation and simplification of arbitrary simplicial complexes, under submission at Computer Graphics Forum.

[17] K. F. Kee, L. Sparks, D. C. Struppa, and M. Mannucci. Social groups, social media, and higher dimensional social structures: a simplicial model of social aggregation for computational communication research. *Communication Quarterly*, 61(1):35–58, 2013.

[18] W.-C. Lee and M. Ye. Location-based social networks. In *Encyclopedia of Social Network Analysis and Mining*, pages 821–833. Springer, 2014.

[19] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[20] R. D. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2):169–190, 1950.

[21] J. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *NIPS*, volume 2012, pages 548–56, 2012.

[22] T. J. Moore, R. J. Drost, P. Basu, R. Ramanathan, and A. Swami. Analyzing collaboration networks using simplicial complexes: a case study. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, pages 238–243. IEEE, 2012.

[23] A. Muhammad and A. Jadbabaie. Decentralized computation of homology groups in networks by gossip. In *American Control Conference*, pages 3438–3443, 2007.

[24] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[25] S. Pemmaraju and S. Skiena. *Computational discrete mathematics: combinatorics and graph theory with mathematica*. Cambridge University Press, New York, NY, USA, 2003.

[26] W. Ren, Q. Zhao, R. Ramanathan, J. Gao, A. Swami, A. Bar-Noy, M. P. Johnson, and P. Basu. Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes. *Wireless networks*, 19(6):1121–1133, 2013.

[27] H. Samet. *Foundations of multidimensional and metric data structures*. The Morgan Kaufmann series in computer graphics and geometric modeling. Morgan Kaufmann, 2006.

[28] J. Scott. *Social network analysis*. Sage, 2012.

[29] P. Symeonidis, D. Ntempos, and Y. Manolopoulos. Location-based social networks. In *Recommender Systems for Location-based Social Networks*, pages 35–48. Springer, 2014.

[30] S. Wasserman and K. Faust. *Social network analysis: methods and applications*, volume 8. Cambridge university press, 1994.

[31] A. C. Wilkerson, T. J. Moore, A. Swami, and H. Krim. Simplifying the homology of networks via strong collapses. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 5258–5262. IEEE, 2013.