A Model for Urban Social Networks

Stefano Guarino, Enrico Mastrostefano, Alessandro Celestini, Massimo Bernaschi, Marco Cianfriglia, Davide Torre, and Lena Rebecca Zastrow

Istituto per le Applicazioni del Calcolo "Mauro Picone",
Consiglio Nazionale delle Ricerche, Rome, Italy
{n.surname}@iac.cnr.it

Abstract. Defining accurate and flexible models for real-world networks of human beings is instrumental to understand the observed properties of phenomena taking place across those networks and to support computer simulations of dynamic processes of interest for several areas of research – including computational epidemiology, which is recently high on the agenda. In this paper we present a flexible model to generate age-stratified and geo-referenced synthetic social networks on the basis of widely available aggregated demographic data and, possibly, of estimated age-based social mixing patterns. Using the Italian city of Florence as a case study, we characterize our network model under selected configurations and we show its potential as a building block for the simulation of infections' propagation. A fully operational and parametric implementation of our model is released as open-source.

Keywords: Urban social network · Graph model · Simulator · Epidemic.

1 Introduction and Background

The definition of networks that encode in a suitable way patterns of connection and interaction among individuals of a population is a widely studied problem. Among the many reasons, finding accurate models for real-world social networks is instrumental to study the dynamics of disease spreading [9] or propaganda [14]. Many simulation-based social studies complain about the lack of reliable data and therefore model social networks by using well-known random graph models [1]. At the other hand of the spectrum, with a focus on physical interactions, a growing body of research makes use of extensive and often purposely collected data – *e.g.*, surveys and questionnaires, activity location, traffic and mobility data – either to extract setting-specific contact matrices [24,3] or for tuning agent-based simulators [10,3]. While simple random models cannot capture all the subtleties of real networks [3], a recent call to action raised the attention towards the need for accurate yet flexible and replicable approaches [32].

In this paper, we present a novel framework for the definition of a data-driven urban social network, where each edge of the graph represents a "strong tie" [18] between two geo-referenced and age-stratified individuals. We tell apart intra-household (*e.g.*, kinship) edges from friendship edges. The former are defined

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

quite naturally by drawing a *clique* (*i.e.*, a complete subgraph) for each household, where the breakdown of the population into households is entirely inferred from the available data. Friendship edges are instead drawn based on three guiding elements: (i) the available contact data (*e.g.*, extracted thanks to [35]), used to trigger an age-based social mixing structure in the network; (ii) the existence of an inverse power-law dependence of friendship upon physical distance [15,7]; (iii) a vertex-intrinsic social fitness [8] that models the individual propensity to have friends. Our network model may be of help in any application setting that requires to gather and elaborate information on the urban social fabric. It may be used as a standalone tool, to characterize urban social relation patterns in connection with the geography and the demography of a given territory – like we show in Section 3. Moreover, it is instrumental in increasing the plausibility of simulations of dynamic processes that may be influenced by agents' preferences and personal relations – like we show in Section 4 for an epidemic use case.

To guarantee usability and reproducibility, the source code of the software used to simulate instances of our urban social network is publicly available[1]. The model depends on a combination of data-driven and configuration parameters that make it adaptable to different use cases. In [13] we provide a detailed analysis that may guide potential users and that, overall, speaks in favor of certain configurations, on which we will focus in this paper. Of special interest are the combinations of parameters that allow to reproduce a few empirical and sociological findings of urban social networks. First of all, the distance-based penalization shall have exponent in the range $[0.5, 2]$ [19,26]. Further, the graph shall have a heavy- but not fat-tailed degree distribution [19,15,20,16] and be (mostly) connected, as typically observed in urban areas [27,15,20,31]. Since social ties comparable to kinship are rare [18], these properties will be enforced while keeping "small" the average number of friends. As a consequence, in our graph acquaintances correspond to short, but > 1, paths, and the network is quite sparse, a necessary feature in most practical applications.

A review of related work follows. We then describe in details our network model (Section 2), characterize the network obtained for the city of Florence, Italy, under two selected configurations (Section 3), and present an epidemic use case (Section 4). Finally, we discuss strengths and current limitations of our model, and we identify suitable directions for future work (Section 5).

1.1 Related Work

We construct our synthetic population following an intermediate approach between *Synthetic Reconstruction* (SR) [5] and *Combinatorial Optimization* (CO) [33]. In SR the attributes of each agent are drawn from joint-distributions deduced from aggregate and survey data. In CO a sample of real individuals is available for different sub-areas of the territory of interest, and the whole population is obtained through replication/resampling methods. Extensions and modifications to these methods are well surveyed in [30].

[1] The source code is released under the GPL v3 at gitlab.com/cranic-group/usn.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

One element of novelty of our network model is the usage of age-based social mixing data to infer *friendship* links. Computational social scientists often rely on rather simple graph models [1] or, possibly, on exponential random graphs [29]. Sample data (*e.g.*, surveys, questionnaires, diaries), possibly integrated with mobile/traffic/wearable sensor data [24,11,17], are extensively used to model physical contacts. To this end, some authors extract *contact matrices* for specific settings, such as households, schools and workplaces [24,3], others use agent-based simulators to reproduce synthetic interactions [10,3]. We relied on the recently released SOCRATES [35] Data Tool[2] to extract data for Italy from Polymod [24]. The tool allows to easily specify parameters such as age breaks, gender, day of the week, duration or location of the contacts, and it produces a social contact matrix drawing from the best public survey datasets for the selected country.

The introduction of a penalization for "long" edges is not peculiar to our model. While there is wide evidence that geographical factors alone cannot explain the structure of real-world spatial social networks [31,20,15], the dependence of friendship on distance is widely assumed to follow an inverse power-law with exponent $\beta \in [0.5, 2]$ [19,31,15,26,20,16,34,7] – and this surprisingly holds even for online relationships [12]. In particular, $\beta < 1$ seems to work better for short range contacts ($< 20km$) [16] and for urban networks [34]. The impact of this penalization upon communities, path lengths, degree distribution and other topological properties of the network has already been the object of study [36,4], but previous modeling efforts assumed some simple (*e.g.*, uniform) spatial distribution, instead of using data-driven vertex locations.

Previous empirical findings did play a role, more generally, in guiding our modeling choices. Real-world spatial social networks are usually "small-worlds" [31,15], with a single giant connected component [27,15,20], average degree in the range 5 to 20, and high clustering coefficient [20,31,15,20,16]. In line with sociological studies [18], but contrary to other real-world networks [6,25], such networks do not present very large hubs [27,19,15,20]. Their degree distribution is right skewed and relatively long-tailed [15,20], and it has been, at times, approximated by a power-law with a large (5 to 8) exponent [27,19] or by a Lognormal distribution [16]. Within cities, population density impacts on the frequency of close-range contacts, but usually not on the overall size of each person's network [7]. While geographical proximity and community structure appear to be related [15,34,7], some authors argue that only small clusters (< 30 members) are geographically bounded [26] whereas the large ones may span across very large areas of a city [15].

2 Graph Model

Our urban social network is represented by an unweighted undirected graph $G = (V, E)$, where V is the vertex set of size $N = |V|$ and E is the edge set. In particular, we have $E = E_H \sqcup E_F$, where E_H is the set of *household edges*, E_F

[2] `https://lwillem.shinyapps.io/socrates_rshiny/`.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

is the set of *friendship edges* and $\sqcup$ denotes the disjoint union. In the following, we explain how V, E_H and E_F are defined in our model. We will often use the expression *household graph* to denote the subgraph $G_H = (V, E_H)$ and the expression *friendship graph* to denote the subgraph $G_F = (V, E_F)$.

2.1 Vertex set

Each vertex $u \in V$ is characterized by three attributes: a fitness score $f_u \geq 0$, an age label $g_u \in \{0, \ldots, n-1\}$, and a tile label $t_u \in \{0, \ldots, T-1\}$.

Fitness. Inspired by previous work, that modelled degree heterogeneity by means of a vertex-intrinsic fitness [8], we make use of a *sociability fitness* attribute f_u. Our model does not put restrictions upon the choice of f_u, but the probability of a friendship edge between u and v is set proportional to f_u and f_v (see Section 2.3). The distribution of f_u shall thus be chosen considering its impact on the degree distribution of the friendship graph. For the scope of this paper, we consider $f_u \sim 1 + \mathcal{LN}\left(\ln(2), \frac{1}{4}\right)$, where $\mathcal{LN}$ denotes a Lognormal distribution[3]. This distribution has been chosen empirically in an attempt to mimic two main aspects of real-world spatial social networks: only a few people have very few social links and the hubs are limited in both number and size. In general, Lognormally distributed data occur across different domains [23] and recent work suggests that the sociability of real-world social networks makes no exception [20,16]. Other choices may be preferred, some of which (*e.g.*, a Pareto, a uniform and a constant distribution) are already supported by our simulator.

Age. The age labels define a stratification of the population into age-groups, *i.e.*, a partition of the vertex set V into n disjoint subsets $V_0, \ldots, V_{n-1}$. For the scope of this paper, we consider four age-groups: *children* (0 to 17), *young* people (18 to 34), *adults* (35 to 64) and *elderly* people (65+). The proportion of each group is determined according to census data at the provincial level made available by the Italian Institute of Statistics (ISTAT)[4] and for each vertex u the age label g_u is independently drawn. Any desired age-stratification can be easily specified in the simulator's configuration file – statistics for many other countries are provided, for instance, by the United Nations Statistics Division (UNSD)[5].

Tile. We decompose the territory of interest into a regular lattice of T square tiles of side l and we set the tile label t_u equal to the unique index of the tile where u resides. The side l is a configuration parameter, set as $l = 1$Km for the scope of this paper. Approximating the position of each vertex with its tile is instrumental in simplifying the computation of pairwise distances and of the household structure, as better explained in the following. A module of the simulator is responsible for extracting the shape file of the city of interest. We resort to the `overpass` API of the well known OpenStreetMap database[6] to

[3] Throughout this paper, we use the parameterization $\mathcal{LN}\left(\lambda, \sigma^2\right)$ where λ and σ^2 are the mean and variance of the associated Normal distribution.

[4] ISTAT data used in this paper are available at `https://www.demo.istat.it/pop2020`

[5] `https://unstats.un.org/unsd/demographic-social/census/censusdates/`.

[6] `https://www.openstreetmap.org/`.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

find the minimal grid that contains the city's boundary; we then select only the tiles of the grid whose center lies inside it. We get population density data for the whole city from the WorldPop Project[7], which provides data of the world population for 100m×100m square cells, and we map those data to our tiles.

2.2 Household edges

To group individuals into households we follow a heuristic approach, imposing that: (*i*) all members of a household live in the same tile; (*ii*) children are younger than their parents; (*iii*) partners have, on average, a similar age. The algorithm is based on the concept of *household role*, represented as a pair of the form (household-type, role) taking values in {(singles,single), (single-parent,parent), (single-parent,child), (couples,peer), (two-parents,parent), (two-parents,child), (various,various)}[8]. For instance, r_u =(single-parent,parent) means that u is a parent in a household of type single-parent, where $r_u[0]$=single-parent is the household-type and $r_u[1]$=parent is the role. We make use of two conditional distributions: $\Pr[r \mid g]$ is the probability that an individual has role r given that she/he belongs to age-group g; $\Pr[k \mid h]$ is the probability that a household of type h has k members. These can be obtained for Italy based on ISTAT aggregate national data, and, *e.g.*, from the UNSD for other countries. At a high level, the heuristics works as follows:

- Extract a role r for each vertex u, based on $\Pr[r \mid g_u]$.
- For all u such that $r_u[0] \in \{\text{single-parent}, \text{two-parents}\}$:
  - if $r_u[0]$ =two-parents, select a random partner v for u such that $t_v = t_u$, $g_v \in [g_u - 1, g_u + 1]$ and $r_v[0] = r_u[0]$;
  - extract the total number of members k_u for the household of u, based on $\Pr[k \mid r_u[0]]$, and compute their total number of children c_u.
- For $i = 1, \ldots, \max_u c_u$:
  - for all u such that $c_u \geq i$, select a random w such that $t_w = t_u$, $g_w < g_u$, $r_w[0] = r_u[0]$ and $r_w[1]$=child, and assign w to the household of u.
- For all u such that $r_u[0]$ =couples, select a random partner v for u such that $t_v = t_u$, $g_v \in [g_u - 1, g_u + 1]$ and $r_v[0] = r_u[0]$.
- Randomly compose the households of type various, based on $\Pr[k \mid \text{various}]$.

In our simulations, the number of individuals left out of any household by the heuristics is negligible, and the empirical distributions of household types and members per type almost perfectly match the expected ones (see [13] for details). The household edges E_H are finally obtained as the union of all the *cliques* that connect all members of the same household.

[7] `https://www.worldpop.org/`.
[8] The pair (various,various) covers all cases other than the previous ones.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

2.3 Friendship edges

All friendship edges are drawn independently at random. For each pair $(u, v) \in V \times V$, the probability $\Pr[u, v] = \Pr[(u, v) \in E_F]$ is defined as:

$$\Pr[u,v] = \frac{\mu \cdot N}{2} \cdot \frac{m_{g_u,g_v} \cdot s_{g_u,g_v}}{\sum_{i \leq j} (m_{i,j} \cdot s_{i,j})} \cdot \frac{d(u,v)^{-\beta} \cdot f_u \cdot f_v}{\sum_{u' \in V_{g_u}, v' \in V_{g_v}} (d(u',v')^{-\beta} \cdot f_{u'} \cdot f_{v'})} \quad (2.1)$$

where:

- μ is the average number of friends – a configuration parameter;
- $m_{i,j} = |V_i| \cdot |V_j|$ if $i \neq j$ and $m_{i,i} = \frac{|V_i| \cdot (|V_i|-1)}{2}$ – deduced from the data-driven age-stratification;
- $s_{i,j}$ is the age-based social mixing for groups i and j – a data-driven coefficient, computed from aggregated social contact data as explained in Section 1.1;
- $d(u, v) = \max\left\{\frac{l}{2}, d^*(t_u, t_v)\right\}$ is the approximated distance between u and v, where $d^*(t_u, t_v)$ is the distance between the centers of the tiles t_u and t_v – a data-driven value, except for l which is a configuration parameter;
- β is the exponent that determines the level of penalty imposed to long edges – a configuration parameter.

A thorough description of (2.1) is presented in [13]. Here, we just highlight that $\Pr[u, v]$ is normalized in such a way to guarantee that the data-driven age-based social mixing induced by the coefficients $s_{i,j}$ is respected, up to a scaling factor. Indeed, the expected number of friendship edges between groups i and j is

$$\mathbf{E}[|E_F(i,j)|] = \frac{\mu \cdot N}{2} \cdot \frac{m_{i,j} \cdot s_{i,j}}{\sum_{i \leq j} (m_{i,j} \cdot s_{i,j})}$$

It follows quite easily that $\mathbf{E}[|E_F|] = \frac{\mu \cdot N}{2}$, hence the average degree of the friendship graph is exactly μ, regardless of all other parameters. The expected degree of a specific vertex u is proportional, besides to μ, to f_u and to the average of f_v for all other $v \in V$, weighted by $d(u, v)^{-\beta}$.

3 Network Analysis

Potential sources of information for real friendship patterns, *e.g.*, telephone data [11] or online social networks [21], are usually hard to acquire, private and/or not entirely representative/dependable. Instead of a direct validation of our model against real data, we therefore present a characterization of the urban social network obtained for the city of Florence under selected configurations. We refer the interested reader to [13] for an extensive experimental analysis.

In the following, we use age-stratification and household composition data from ISTAT, spatial population density from WorldPop, and age-based social mixing coefficients from [24], collected through the SOCRATES Data Tool. We additionally take $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$, $\mu = 10$ and we consider both $\beta = 0.5$ and $\beta = 2$. It may be useful to know that, based on our data, Florence counts 363060 residents – roughly, 15% children, 17% young people, 43% adults, 25% elderly people – and is contained in a 15Km×12Km grid.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

3.1 Topology of the graph

In Table 1, we overview the global topological properties of the graph. We recognize the typical positive assortativity of social networks and a global clustering several orders of magnitude greater than in the equivalent Erdos-Renyi graph. However, a closer inspection highlights that the large number of small cliques introduced in the household graph plays a paramount role in the formation of triangles, whereas the friendship graph, despite the geographical and age-based homophily, shows limited transitivity. Regardless of β, the average shortest path length has a value of the order $\frac{\ln(N)}{\ln(\langle \text{deg} \rangle)}$, typical of small world networks.

Table 1: Social graph for Florence with $\mu = 10$ and $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$: average metrics over 10 independent runs (the negligible variance is omitted).

	$\langle \text{deg} \rangle$	$\langle \text{dist} \rangle$	C	C_{loc}	ρ	# comp.	giant %
$\beta = 0.5$	11.812	5.2633	0.0156	0.0325	0.2106	924.9	99.74%
$\beta = 2$	11.815	5.3199	0.0148	0.0438	0.2605	2333.1	99.28%

$\langle \text{deg} \rangle$: average degree; $\langle \text{dist} \rangle$: average path length; C: global clustering coefficient; C_{loc}: average local clustering coefficient; ρ: degree assortativity; # comp.: number of connected components; giant %: percentage of nodes in the giant component.

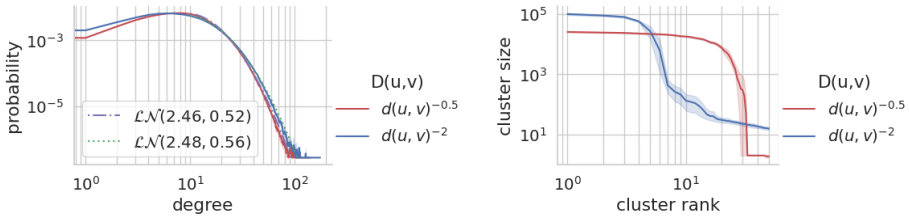
From Fig. 1a we see that, as expected, the right tail of the degree distribution is heavy but not fat (*i.e.*, subexponential but not power-law) – as a matter of fact, the frequency of degrees $\geq \mu$ is well-fitted by a Lognormal distribution. Comparing the two regimes for β, we see that $\beta = 2$ yields a larger portion of loosely connected vertices compensated by the presence of greater hubs. The rationale is that only when the dependence on the distance is weak the individuals living in central and denser areas connect to peripheral vertices, that remain otherwise isolated. $\beta = 2$ thus favors the assortativity and the average local clustering, but causes a greater number of connected components. In any case, the giant component consistently covers more than 99% of the graph.

For what concerns the organization of our network in communities, we consider modularity-based clusters obtained with the Louvain algorithm [25]. From Fig. 1b, we see that when $\beta = 0.5$ the network *de facto* consists of ≈ 20 clusters of comparable size. The relatively low modularity of the obtained partition (≈ 0.27) indicates that these clusters are significantly intertwined. Conversely, when $\beta = 2$ most nodes of the network lie in few well-defined giant communities, surrounded by a multitude of communities of variable size.

3.2 Socio-geography of the graph

Since our model incorporates a penalization for long edges, we expect some indication of correlation between topological properties and population density.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

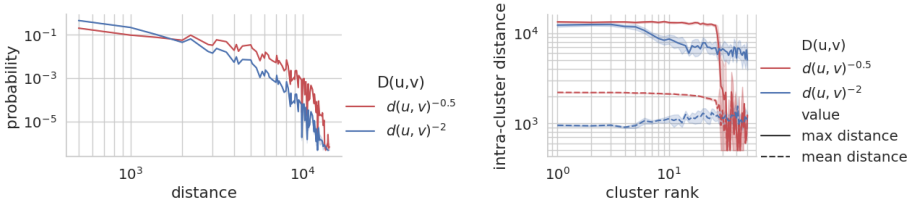


(a) Degree distribution with $\mathcal{LN}$ fit. (b) Size of the largest 50 communities.

Fig. 1: Social graph for Florence with $\mu = 10$ and $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$: average features with a 95% confidence interval over 10 independent runs.

The first, almost obvious, finding is that setting $\beta = 0.5$ significantly favors the creation of long edges at the expenses of very short ones, as shown in Fig. 2a. Notably, the distribution of the edges' physical length does not depend on the chosen μ and f_u, but it is entirely controlled by β.

In Fig. 2b we show the mean and max intra-cluster distance for the first 50 clusters of the graph. In line with empirical findings [26], only very small communities are geographically bounded – this is especially visible for $\beta = 0.5$ due to the sudden drop in community size emerged in Figure 1b. Remarkably, when $\beta = 2$ the mean intra-cluster distance is often comparable to the tile side l (set to 1Km as per Section 2.1), meaning that, even in large clusters, most adjacent vertices are at one tile of distance or less. When $\beta = 0.5$, instead, the mean distance consistently lies between $2l$ and $3l$.



(a) Physical distance of adjacent vertices. (b) Mean/max intra-cluster distance.

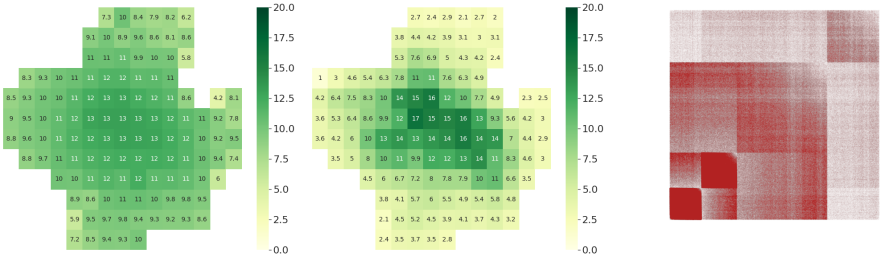
Fig. 2: Social graph for Florence with $\mu = 10$ and $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$: average features with a 95% confidence interval over 10 independent runs.

It is reasonable to expect that vertices that are closer, on average, to other vertices will generally have a greater degree. This is confirmed by Figs. 3a and 3b, two heatmaps where the color gradient indicates the average degree of each tile. In particular, with $\beta = 2$, most tiles are far below average whereas the tiles surrounded by a densely populated area have a high average degree. The introduction of a social fitness attribute makes it possible to achieve the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

heterogeneity of sociable individuals within each tile. Yet, on average, the vertices having a favorable position in the territory will have a greater degree and the main hubs will be individuals with large f_u living in densely populated areas.

Finally, in Figure 3c we plot the graph's adjacency matrix for $\beta = 0.5$ (the case $\beta = 2$ being completely alike), where nodes are ordered by their age-group. The observable assortativity by age, inherited by the data-driven coefficients $s_{i,j}$, is clear and in qualitative agreement with previous work on social mixing patterns [10,17,28,22]. In analogous contact matrices, it is often possible to identify sub-diagonals which account for parent-children contacts [10,22]. Such sub-diagonals are, in our case, non-detectable having just four age-groups.



(a) Average degree of each tile for $\beta = 0.5$.

(b) Average degree of each tile for $\beta = 2$.

(c) Adjacency matrix for $\beta = 0.5$, nodes sorted by age.

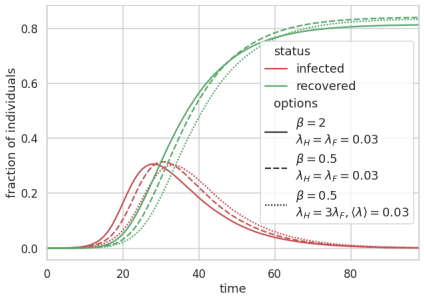
Fig. 3: Social graph for Florence with $\mu = 10$ and $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$.

4 Epidemic Use Case

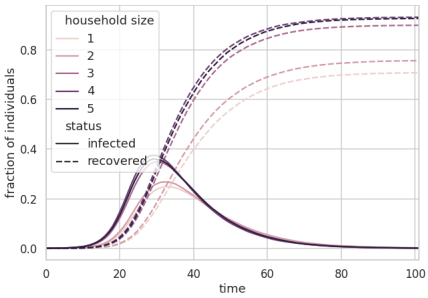
To further assess the practical relevance of our model, we simulated a SIR epidemic upon the giant connected component of the synthetic graphs obtained with $\mu = 10$, $f_u \sim 1 + \mathcal{LN}(\ln(2), 0.25)$ and, unless otherwise specified, $\beta = 0.5$. We consider a discrete-time *synchronous cellular automaton* in which the dynamic follows a *reactive process* [2]: at each time step, each infected individual spreads the disease to each of its neighbors with probability λ and recovers with probability δ. For the scope of this use case, we arbitrarily set $\delta = 0.1$ and $\lambda = 0.03$. If I_t denotes the set of infected individuals at time t, we assume that $|I_0| = 100$, *i.e.*, $< 0.03\%$ of the population is infected at time 0. Albeit typical epidemic simulations consider possibly dynamic and denser networks of contacts, our network of strong ties may be interpreted as a coarse-grained model for highly-infectious, frequent and close contacts. In the following, we aim at showing that the parametric and data-driven nature of our model allows to draw high-level indications about the impact of several socio-demographic and geographic features on the epidemic.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

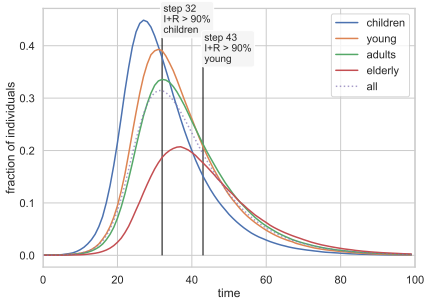
In Fig. 4a we show the evolution of the fraction of infected and recovered individuals for different combinations of β and λ. When $\beta = 2$, the higher frequency of edges in densely populated areas favors a quicker spread of the infection, but the existence of loosely connected areas makes the total number of infected nodes slightly lower with respect to the case $\beta = 0.5$. We also see that, if household edges are three times more likely to transmit the disease than friendship edges (*i.e.*, $\lambda_H = 3\lambda_F$), but the overall average infection probability is still $\langle\lambda\rangle = 0.03$, the epidemic is a bit slower but, eventually, equally pervasive. As shown in Fig. 4b for $\lambda_F = \lambda_H = 0.03$, people living in households of size ≥ 3 have a significantly greater chance of catching the infection, probably due to the combined effect of having, on average, a greater degree and of the presence of children and young people in the household.



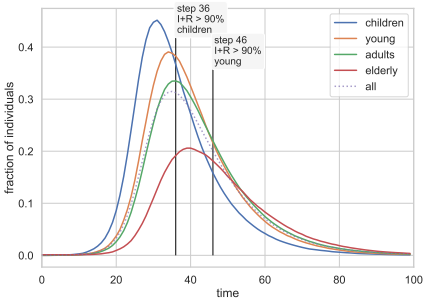
(a) SIR evolution as β and λ vary.



(b) SIR evolution by household size.



(c) I_t by age-group, with I_0 chosen uniformly at random.



(d) I_t by age-group, with I_0 chosen among elderly people only.

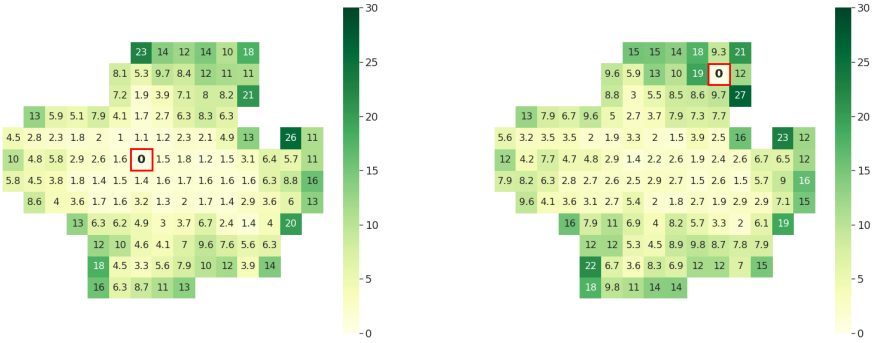
Fig. 4: Evolution of the fraction of infected and recovered individuals for different system parameters, within different households and within different age-groups.

Since our model incorporates a data-driven age-based social mixing, it naturally lends itself to an analysis of the evolution of the epidemic inside single age-groups. From Figs. 4c and 4d, we see that children and young people experience a higher and earlier peak, and they are the only age-groups that reach a 90% prevalence of infected individuals. The younger groups are the drivers of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

the epidemic and the most infected, despite not being the largest groups, but probably due to their strong internal cohesion. In contrast with Fig. 4c, where the individuals in I_0 are chosen uniformly at random, in Fig. 4d all 100 individuals of I_0 are elderly people. A bit surprisingly, in this scenario we only notice a time shift in the epidemic, suggesting that the qualitative behavior of the epidemic depends on when the contagion reaches the younger individuals.

In Fig. 5 we consider the average time, over 10 independent runs, of the first infection occurring in each tile. The whole city center is reached in just a few days both if the infection starts from a central and densely populated tile (Fig. 5a) and if it starts from a peripheral and sparsely populated tile (Fig. 5b). Yet, some areas may be preserved if isolated within one or even two weeks. The starting position of the infection does play a role in our model, with an approximate 50% delay in the time of the first infection for most tiles if the infection starts in the periphery. In that case, the epidemic does not propagate locally but, apparently, it reaches the center before moving back outskirt.



(a) I_0 chosen in a central and densely populated tile.

(b) I_0 chosen in a peripheral and sparsely populated tile.

Fig. 5: Time step of the first infected individual per tile.

5 Discussion and Conclusions

We have implemented a probabilistic model that mimics the strong social ties among a set of nodes representing the population of a given territory, organized into households. Our model is based on just a few clear assumptions: (*i*) not all individuals are equally sociable; (*ii*) the geographical distance and the age difference play a role in the probability that two individuals become friends; (*iii*) it is often fundamental to rely only on data that is already widely available. The simulator provides a way to recreate synthetic social networks within an arbitrary territory for which the social mixing patterns can be inherited from any already existing dataset, thus addressing the common circumstance where aggregated

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

demographic data and some estimate of the age mixing patterns are the only available information. Since the evolution of a social network is significantly slower than most processes of interest occurring on it, our model is, by construction, static. Exploring possible approaches to generate a dynamic interaction network on top of our static social network is among the first directions for future work.

We evaluated our urban social network for the city of Florence, focusing on two configurations selected in light of previous empirical findings – for a more detailed analysis of the model we refer the reader to [13]. With only 10 friends on average, the giant component spans more than 99% of the network. Age and proximity based homophily guarantees the intended internal cohesion of single age-groups and a positive assortativity, yet the transitivity remains weak. By introducing a Lognormally distributed sociability we obtain the often desired heavy-tailed degree distribution. However, when long edges are strongly penalized, sociable hubs tend to concentrate in densely populated areas, thus intensifying the correlation between favorable positioning and degree. Nevertheless, the variability of the degree internal to a tile is preserved, being entirely controlled by the social fitness. A weak penalization of long edges, on the other hand, makes it more difficult to partition the network into well-defined communities. Almost regardless of their size, the communities tend to have a large spatial extension, even though the average distance of their members is small.

Some of the above properties are reflected in the outcomes of a SIR epidemic simulation on the network. The penalization imposed to long edges has an impact on the speed and pervasiveness – both quantitatively and geographically – of the contagion, whereas the age-based social mixing determines which age-groups drive the infection to a greater extent than the prevalence of different age-groups. Regardless of the specific epidemic use case considered in this paper, our model appears well suited to support the analysis of dynamic processes occurring within a urban population, thanks to its adherence to real data and its flexibility that allow to easily evaluate the impact of socio-demographic and geographic features.

The intrinsic ambiguity in the concept of "friendship" leaves a few issues, somehow, open to further investigations. We plan to explore the integration of explicit preferential attachment mechanisms, and to verify whether a fine-grained age-stratification or age-specific β's do foster triangles within certain (*e.g.*, school age) groups. Further, we will consider a density-aware dependence on the distance [20], to gain more control on the social advantage associated to high density areas – which may still be desirable if the goal is predicting physical interactions, *e.g.*, for use in computational epidemiology. Finally, closed results binding the distribution of the social fitness to the obtained degree distribution may significantly improve the usability of our simulator. That said, we believe that our model presents unique features that make it a valuable resource for computational social scientists. Since the simulator is fully parametric and available as open source software, any potential user may adjust it to her/his needs and possibly contribute to its further development.

Acknowledgment. The authors thank the municipality of Florence for the kind support provided and Francesca Colaiori for useful discussions.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

References

1. Amblard, F., Bouadjio-Boulic, A., Gutiérrez, C.S., Gaudou, B.: Which models are used in social simulation to generate social networks? a review of 17 years of publications in jasss. In: 2015 Winter Simulation Conference (WSC). pp. 4021–4032. IEEE (2015)
2. de Arruda, G.F., Rodrigues, F.A., Moreno, Y.: Fundamentals of spreading processes in single and multilayer complex networks. Physics Reports **756**, 1–59 (2018)
3. Barrett, C.L., Beckman, R.J., Khan, M., Kumar, V.A., Marathe, M.V., Stretz, P.E., Dutta, T., Lewis, B.: Generation and analysis of large synthetic social contact networks. In: Proceedings of the 2009 Winter Simulation Conference (WSC). pp. 1003–1014. IEEE (2009)
4. Barthélemy, M.: Spatial networks. Physics Reports **499**(1-3), 1–101 (2011)
5. Beckman, R.J., Baggerly, K.A., McKay, M.D.: Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice **30**(6), 415–429 (1996)
6. Bernaschi, M., Celestini, A., Guarino, S., Lombardi, F., Mastrostefano, E.: Spiders like onions: on the network of tor hidden services. In: The World Wide Web Conference. pp. 105–115 (2019)
7. Büchel, K., Ehrlich, M.V.: Cities and the structure of social interactions: Evidence from mobile phone data. Journal of Urban Economics **119**, 103276 (2020)
8. Caldarelli, G., Capocci, A., De Los Rios, P., Munoz, M.A.: Scale-free networks from varying vertex intrinsic fitness. Physical review letters **89**(25), 258702 (2002)
9. Cauchemez, S., Bhattarai, A., Marchbanks, T.L., Fagan, R.P., Ostroff, S., Ferguson, N.M., Swerdlow, D.a.: Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza. Proceedings of the National Academy of Sciences **108**(7), 2825–2830 (2011)
10. Del Valle, S.Y., Hyman, J.M., Hethcote, H.W., Eubank, S.G.: Mixing patterns between age groups in social networks. Social Networks **29**(4), 539–554 (2007)
11. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences **106**(36), 15274–15278 (2009)
12. Goldenberg, J., Levy, M.: Distance is not dead: Social interaction and geographical distance in the internet era. arXiv:0906.3202 (2009)
13. Guarino, S., Mastrostefano, E., Bernaschi, M., Celestini, A., Cianfriglia, M., Torre, D., Zastrow, L.: Inferring urban social networks from publicly available data. Future Internet (in press) (2021)
14. Guarino, S., Trino, N., Celestini, A., Chessa, A., Riotta, G.: Characterizing networks of propaganda on twitter: a case study. Applied Network Science **5**(1), 1–22 (2020)
15. Herrera-Yagüe, C., Schneider, C.M., Couronné, T., Smoreda, Z., Benito, R.M., Zufiria, P.J., Gonzalez, M.C.: The anatomy of urban social networks and its implications in the searchability problem. Scientific reports **5**, 10265 (2015)
16. Illenberger, J., Nagel, K., Flötteröd, G.: The role of spatial interaction in social networks. Networks and Spatial Economics **13**(3), 255–282 (2013)
17. Klepac, P., Kucharski, A.J., Conlan, A.J., Kissler, S., Tang, M., Fry, H., Gog, J.R.: Contacts in context: large-scale setting-specific social mixing matrices from the bbc pandemic project. medRxiv (2020)
18. Krackhardt, D., Nohria, N., Eccles, B.: The strength of strong ties. Networks in the knowledge economy **82** (2003)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23

19. Lambiotte, R., Blondel, V.D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P.: Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications **387**(21), 5317–5325 (2008)
20. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences **102**(33), 11623–11628 (2005)
21. Mastrandrea, R., Fournet, J., Barrat, A.: Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. PLOS ONE **10**(9), 1–26 (09 2015)
22. Mistry, D., Litvinova, M., Chinazzi, M., Fumanelli, L., Gomes, M.F., Haque, S.A., Liu, Q.H., Mu, K., Xiong, X., Halloran, M.E., et al.: Inferring high-resolution human mixing patterns for disease modeling. arXiv:2003.01214 (2020)
23. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet mathematics **1**(2), 226–251 (2004)
24. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., Edmunds, W.J.: Social contacts and mixing patterns relevant to the spread of infectious diseases. PLOS Medicine **5**(3), 1–1 (03 2008)
25. Newman, M.: Networks: An Introduction. OUP Oxford (2010)
26. Onnela, J.P., Arbesman, S., González, M.C., Barabási, A.L., Christakis, N.A.: Geographic constraints on social network groups. PLoS one **6**(4), e16939 (2011)
27. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., De Menezes, M.A., Kaski, K., Barabási, A.L., Kertész, J.: Analysis of a large-scale weighted network of one-to-one human communication. New journal of physics **9**(6), 179 (2007)
28. Read, J.M., Lessler, J., Riley, S., Wang, S., Tan, L.J., Kwok, K.O., Guan, Y., Jiang, C.Q., Cummings, D.A.T.: Social mixing patterns in rural and urban areas of southern china. Proceedings of the Royal Society B: Biological Sciences **281**(1785), 20140268 (2014)
29. Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P.: Recent developments in exponential random graph (p*) models for social networks. Social networks **29**(2), 192–215 (2007)
30. Ryan, J., Maoh, H., Kanaroglou, P.: Population synthesis: Comparing the major techniques using a small, complete population of firms. Geographical Analysis **41**(2), 181–203 (2009)
31. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. ICWSM **11**, 329–336 (2011)
32. Squazzoni, F., Polhill, J.G., Edmonds, B., Ahrweiler, P., Antosz, P., Scholz, G., Chappin, E., Borit, M., Verhagen, H., Giardini, F., Gilbert, N.: Computational models that matter during a global pandemic outbreak: A call to action. Journal of Artificial Societies and Social Simulation **23**(2), 10 (2020)
33. Voas, D., Williamson, P.: An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography **6**, 349–366 (09 2000)
34. Walsh, F., Pozdnoukhov, A.: Spatial structure and dynamics of urban communities (2011)
35. Willem, L., Van Hoang, T., Funk, S., Coletti, P., Beutels, P., Hens, N.: Socrates: An online tool leveraging a social contact data sharing initiative to assess mitigation strategies for covid-19. medRxiv (2020)
36. Wong, L.H., Pattison, P., Robins, G.: A spatial model for social networks. Physica A: Statistical Mechanics and its Applications **360**(1), 99–120 (2006)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_23