Generating fine-grained surrogate temporal networks

Longa, A.; Cencetti, G.; Lehmann, S.; Passerini, A.; Lepri, B.

Published in:
Communications Physics

Link to article, DOI:
10.1038/s42005-023-01517-1

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

communications physics

ARTICLE

https://doi.org/10.1038/s42005-023-01517-1 OPEN

Generating fine-grained surrogate temporal networks

A. Longa [1,2], G. Cencetti[1,3], S. Lehmann [4,5], A. Passerini [2] & B. Lepri [1✉]

Temporal networks are essential for modeling and understanding time-dependent systems, from social interactions to biological systems. However, real-world data to construct meaningful temporal networks are expensive to collect or unshareable due to privacy concerns. Generating arbitrarily large and anonymized synthetic graphs with the properties of real-world networks, namely surrogate networks, is a potential way to bypass the problem. However, it is not easy to build surrogate temporal networks which do not lack information on the temporal and/or topological properties of the input network and their correlations. Here, we propose a simple and efficient method that decomposes the input network into star-like structures evolving in time, used in turn to generate a surrogate temporal network. The model is compared with state-of-the-art models in terms of similarity of the generated networks with the original ones, showing its effectiveness and its efficiency in terms of execution time. The simplicity of the algorithm makes it interpretable, extendable and scalable.

[1] Fondazione Bruno Kessler, Trento, Italy. [2] University of Trento, Trento, Italy. [3] Aix-Marseille Univ, Université de Toulon, CNRS, Marseille, France. [4] Tecnical University of Denmark, Kongens Lyngby, Denmark. [5] Copenhagen Center for Social Data Science, Copenhagen, Denmark. ✉email: lepri@fbk.eu

In the past decade, temporal networks have driven breakthroughs in real world systems across biology, communications, social interactions, and mobility. One of the main advantages of temporal networks resides in their ability to capture complex dynamics such as, for instance, diffusion and contagion[1–10]. Here we assume that a temporal network is represented in discrete time with each time step corresponding to a static graph, also referred to as a layer of the network. In order to model realistic dynamics, it is often necessary to employ large temporal networks, including a large number of nodes and long time intervals, i.e. many temporal layers[11–13]. Many state-of-the-art temporal datasets, however, are limited both in the number of nodes and in the number of temporal layers[6,14–17]. When the available data are insufficient – e.g. to simulate long-term effects of epidemics – datasets are extended by simply repeating the same temporal sequence multiple times, a procedure which is known to result in biases[15]. An appealing solution to the problem of insufficient data is to use surrogate temporal networks[18]. Surrogate temporal networks are synthetic datasets which mimic the real-world temporal patterns relevant for a desired use-case. Real networks are indeed known to be characterized by typical patterns of interactions, different in different domains (social, biological, infrastructural, etc.), which can be often recognized and delineated[19,20] and the role of surrogate networks is to try to reproduce them. The surrogates can be designed to involve the desired number of nodes and number of temporal layers, where the actual dynamics are known through smaller studies or via available small datasets. Moreover, in the case of privacy sensitive data, such as fine-grained records of social interactions[21], surrogate data can be generated so as to be freely shareable. Over the past years, a large number of successful algorithms for static network generation have been proposed[22,23]; however, extending these models to the dynamic regime has proven prohibitively difficult, due to greatly increased complexity introduced by the temporal dimension.

Indeed, it has become clear that temporal networks are characterized by a highly non-trivial interplay between the instantaneous network topology at a given time (adjacency, degree distribution, clustering, etc.) and the temporal activation of nodes and links – how each connection changes over time (duration of interactions, patterns by which new links appear and old ones disappear, etc.). From the perspective of an individual node, these two dimensions imply that models must take into account (*i*) time, i.e. the history of what has occurred in the preceding timesteps and (*ii*) instantaneous local topology, i.e. the current activation of the neighboring nodes. The scientific literature is full of studies focusing on the spatial dimension but unable to take into account possible temporal correlations[24–29], or – alternatively – works dedicated to model the behavior of individual nodes in time (for example activity driven models[7,30]) which do not aim to reproduce realistic network topologies[31]. There exist models for link prediction that try to combine temporal and topological dimensions by using small local temporal patterns[32] or building over a backbone of significant links[18]. However, there is currently a dearth of models for generating surrogate networks from scratch that are able to take into account the two dimensions simultaneously. The few works, that do this rely on temporal motifs, like *Dymond*[33] and *STM* (Structural Temporal Modeling)[34], or on deep learning like *TagGen*[35]. These three models described in detail in Methods, represent the state-of-the-art. All these models however suffer from some limitations and some of the characteristics of the original networks are not always well reproduced by the surrogate networks.

In this work we propose an alternative method that is particularly efficient in reproducing temporal networks characterized by high temporal resolution and we test it with a wide range of topological and dynamical measures, comparing it with the above cited approaches.

The method that we propose is based on temporal motifs that are defined with an egocentric perspective. Conceptually, we collect the local interactions of each node for a small number of time steps in a real network, the egocentric temporal neighborhood, and we consider them as representative of that network of interactions. We then use them as building blocks to generate a new synthetic network.

A major advantage of the egocentric perspective (that ignores connections among neighbors of an ego node) is that it allows us to linearize the concept of node neighborhood sidestepping the subgraph isomorphism problem[36], that often represents a bottleneck for the algorithms based on motifs. This makes the generation process fast and scalable both in terms of the number of nodes and the number of temporal snapshots. Speed turns out to be a fundamental feature, because the other existing methods rely on algorithms of considerably higher complexity that prevent those methods from scaling to even moderately-sized networks.

We test the method, named Egocentric Temporal Neighborhood Generator (*ETN-gen*), on a range of different temporal networks. In our testing we mainly use social interactions datasets because of richness and availability of these datasets, but the method is general and can be used to generate any kind of graph. Our results show that the surrogate networks that we generate reflect many of the original networks properties with a high degree of accuracy, not just in terms of specific nodes features, as one might anticipate from the local generating mechanism, but with respect to general features, such as the number of interactions, the number of interacting individuals in time and density of their connections. We notice that the characteristics that are better preserved coincide with the ones that depend on time and describe the temporal behavior of the specific nodes, while global features that deal with the spatial organization of the network and which can be observed for instance when collapsing the temporal layers (like the existence of communities) are more difficult to reproduce with this method.

In general, this work allows us to investigate and set the spatio-temporal scale of the minimal fundamental knowledge that is necessary to capture many of the intrinsic characteristics that we aim to reproduce in a temporal network. The possibility to generate surrogate networks that resemble an original one serves as a test to demonstrate the method efficiency.

Results

We first briefly sketch the temporal graph generation process. Then, we use our method to generate temporal graphs which reproduce the temporal interaction patterns of a diverse set of face-to-face interaction networks, including a hospital[37], a workplace[38], and a high school[39]. See Methods for details on the datasets. We evaluate the quality of the generated networks in terms of interaction statistics, considering both static and temporal network properties, highlighting the advantage of our proposed method relative to the state-of-the-art. Finally, we show how the approach can be used to expand existing temporal networks, both in time and in number of nodes, something which is not possible using other methods for surrogate networks.

The neighborhood generation process. Figure 1 shows a graphical representation of the generation process for a small temporal network with three timesteps (see Methods for details). A formalization, in terms of pseudo-code, of the generation process is provided in Supplementary Note 1. Our generative algorithm uses as building block the Egocentric Temporal Neighborhood[40] of a node ($\mathcal{E}_i^{\{t-k,\ldots,t\}}$ for node i), which represents the

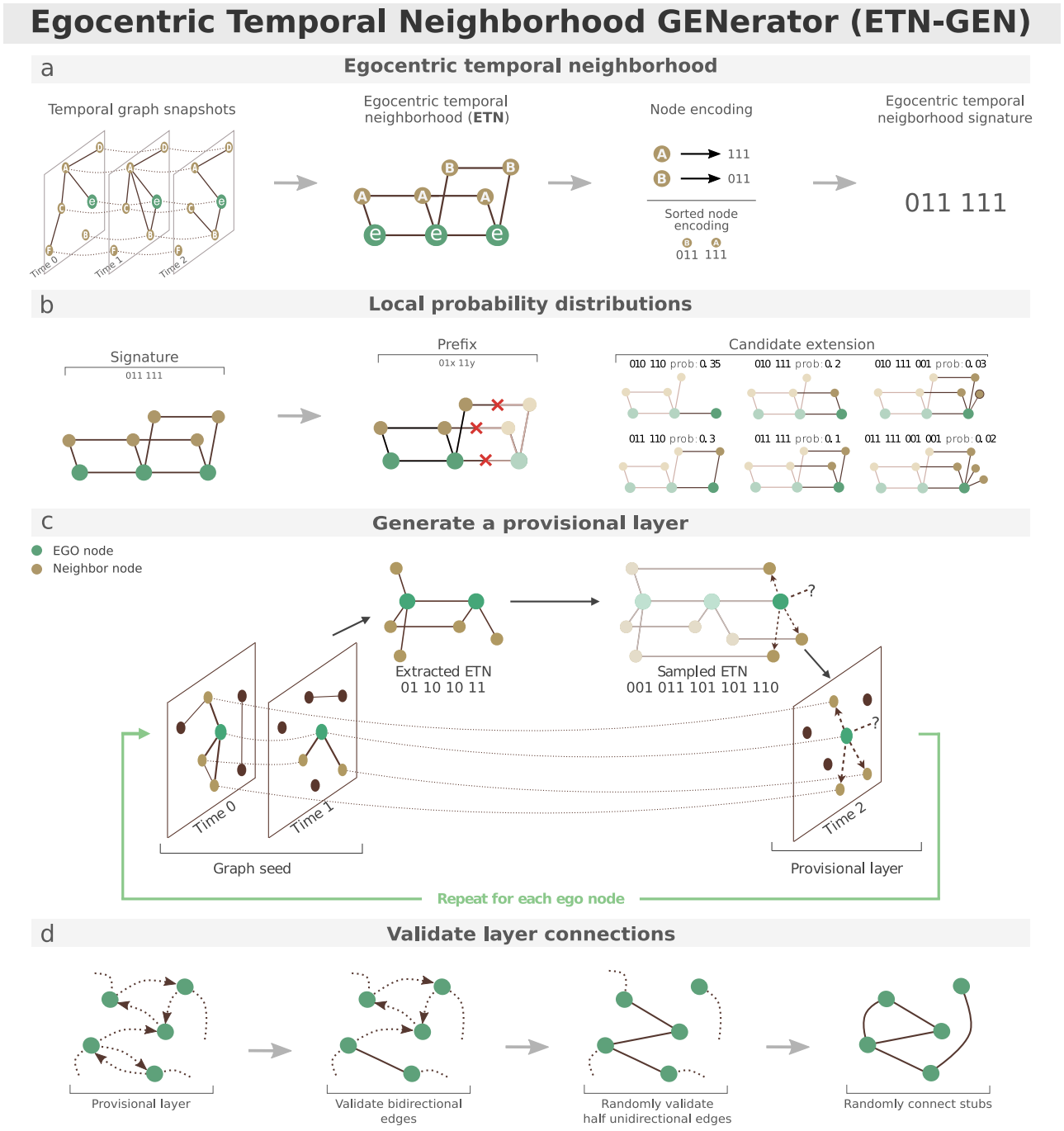

Fig. 1 Egocentric Temporal Neighborhood Generator (*ETN-gen*). Panel **a** shows how egocentric temporal neighborhood signatures are extracted and computed. Panel **b** shows how to build the probability distribution of neighborhoods, necessary to generate a provisional layer. Panel **c** shows how to generate a provisional layer, while panel **d** explains how to convert the provisional layer into a definitive one. Green nodes represent ego nodes, while brown nodes depict neighboring nodes.

neighborhood of a node over a (short) temporal span k. For the sake of compactness, we will refer to the Egocentric Temporal Neighborhood as ETN or simply neighborhood in the rest of the manuscript. Panel a of Fig. 1 shows the neighborhood of a specific node, denoted as e, for a temporal span $k = 2$. $\mathcal{E}_e^{\{t-k,\ldots,t\}}$ contains e and its neighbors at each of $k+1$ consecutive timestamps, discarding connections between the neighbors of e, and adding (temporal) connections among instances of the same node at different timestamps.

Having discarded links between neighbors, $\mathcal{E}_e^{\{t-k,\ldots,t\}}$ can be encoded as a binary string, where for each neighbor node and timestamp 1 (resp. 0) indicates the presence (resp. absence) of a link connecting to the node at that timestamp. Such neighborhoods are extracted for all nodes and all timestamps by using a sliding window over time. Notice that a string of 0 and 1 of length $k+1$ is obtained for each neighbor of e but the identity of these nodes is not stored and the final signature does not include any identity labels, just the shape of interactions between neighbors and e, as shown in the last step of panel A. This implies that the same specific $\mathcal{E}_i^{\{t-k,\ldots,t\}}$ can be found multiple times in the network, referred to different nodes i, different neighbors, and different t.

Second (panel b), we build a local probability distribution designed to enable simulation of activity in future time steps. This distribution to extend the graph into future time steps is based on past neighborhood activity. Specifically the local distribution maps neighborhoods of length $k-1$ (i.e. temporal neighborhoods involving k steps, denoted as Prefix in the figure) to the set of all possible extensions into the future (i.e. neighborhoods of temporal depth k, involving $k+1$ steps), with associated probabilities estimated by Maximum Likelihood over the whole dataset. Basically frequencies of neighborhoods of temporal depth k are first collected from the original temporal network, and then normalized by dividing them by the sum of the frequencies of the neighborhoods sharing the same $k-1$ prefix. These are denoted as Candidate extensions in the figure, where some examples of possible extensions of the prefix are depicted, with their signature and their probability.

Third (panel c), we build the surrogate network layer after layer. Given the first k layers, we generate the subsequent one by sampling future interactions for each node from the local probability distribution described above, thus generating a provisional temporal extension of each node. Notice that since local probability distributions ignore node identities, future interactions can only involve previously existing neighbors or novel (still unknown) nodes (question mark in panel c). All the interactions extracted for each node e are represented as directed links from e to its desired neighbors (including stubs, representing links-to-be to unknown nodes). We thus obtain a provisional directed temporal layer of the network.

Last (panel d), this provisional layer is finalized by combining provisional temporal extensions of all nodes, resolving conflicts and dangling links so as to preserve as much as possible each node's desired neighborhood. We consider a connection from node i to node j in the provisional layer a request of i to be connected to j. If this request is reciprocal the link is validated and added to the new temporal layer (second step in panel d). All remaining one-directional links are validated with probability $\alpha = 1/2$ (third step), to preserve the overall number of connections (an $i-j$ connection can be requested by i or by j). Finally, stubs are pairwise matched up at random (last step in panel d). The procedure is repeated as many times as the desired length of the final temporal graph, always considering the last k timestamps as seeds to generate an additional one.

With the basic mechanisms in place, we take a step back and explain how to initialize the process, i.e. how to obtain the first k layers of the graph. The graph at the first timestamp is generated using a configuration model[41,42] reproducing the degree distribution of the first layer of the original graph. The following layers up to k are generated by applying the procedure in Fig. 1 to the first layer with $k' = 1$, to the first two layers with $k' = 2$ and so on until $k' = k$.

Temporal networks are often characterized by an intrinsic periodicity[1]. This can be captured in our generation process by collecting multiple local probability distributions from the original graph, associated for instance to different days of the week or times of the day. In the experiments in this paper we use distinct week/weekends or daily local probability distributions, depending on the length and variability of the input network.

The recursive procedure poses no limit to the temporal extension of the network, allowing to generate as many temporal layers as desired, even more than those existing in the original network. Plus, the number of nodes too can be set independently of the original network size (the section titled Dataset expansion and extension).

Above, we have described the simplest possible strategy for extending a layer into the future, but note that all random choices in the link validation process could become preferential choices in order to optimize a specific characteristic of the final network (see Section Topological similarity evaluation).

Model evaluation. We now evaluate the quality of the generated networks based on interaction statistics by comparing the networks to empirical data as well as networks generated by a suite of state-of-the-art temporal network generation methods described below. We evaluate performance in terms of individual layer topology as well as temporal behavior.

The state-of-the-art methods we consider are: *Dymond*[33], a model which uses the distribution of 3-nodes structures in the original graph (triads with one, two or three connections) as building blocks to generate a new temporal network; *STM* (Structural Temporal Modeling)[34], a generative model based on the distribution of small temporal motifs; and *TagGen*[35], based on deep learning, which uses a generative adversarial network to generate temporal walks that are then combined into a temporal graph. *Dymond* and *STM* only consider local information, while *TagGen* is more global. These three methods were selected based on the capacity to generate surrogate temporal (rather than static) networks, their performance and the implementation availability.

It is important to underscore that these network generation methods have not necessarily been developed with the aim of generating large temporal networks with low computational cost (see Methods, subsection time and complexity). This means that, for example, they require much more training data, need denser temporal snapshots, and therefore struggle to generate high temporal resolution networks. In particular, both *Dymond* and *STM* require a triad motif to appear in the snapshot. This assumption is too strong for fine-grained snapshots, (i.e. a snapshot every minute). On the other hand, since *TagGen* is based on a deep learning technique, it requires a massive amount of data in the training phase. Differently *ETN-gen*, thanks to the linearization allowed by the egocentric perspective, can scale to arbitrarily sized temporal networks. In the rest of the paper we report experiments only on the three smallest face-to-face interactions datasets, collected in the hospital[37], in the workplace[38], and in one of the high schools[39] respectively. Results applying *ETN-gen* to larger datasets are reported in Supplementary Note 2, 3 and 4 and compared only to *TagGen* due to computational complexity of the other methods.

In Supplementary Note 5, we also show the comparison with two other alternative models which are not specifically designed for surrogate temporal network generation, namely *ADN* (Activity-driven time varying Network Model)[30], a popular method to generate temporal networks, and *CTGAN* (Conditional Tabular GAN)[43], a machine learning approach to generate tables.

Figure 2 reports the total number of interactions for each temporal snapshot (left) and the average number of nodes (right) in the original network, *ETN-gen* and the three state-of-the-art methods. The first clear finding from this figure is that *ETN-gen* (orange curves) results in time-series that are remarkably similar to those appearing in the original datasets (black curves). This is true, not just in terms of generating a number of interactions which is of the same order of magnitude as the original data (notice that different datasets have different scales on the y-axis), but also in terms of temporal patterns which are preserved with considerable accuracy, including daily and weekly periodicity.

This result should not come as a surprise, as it is a direct consequence of our network generation procedure. The local probabilistic models store the probability distributions of the neighborhoods appearing in the original graph and this indirectly contains the key information about how nodes degree evolves in time. Further, our seed-network has the same degree distribution

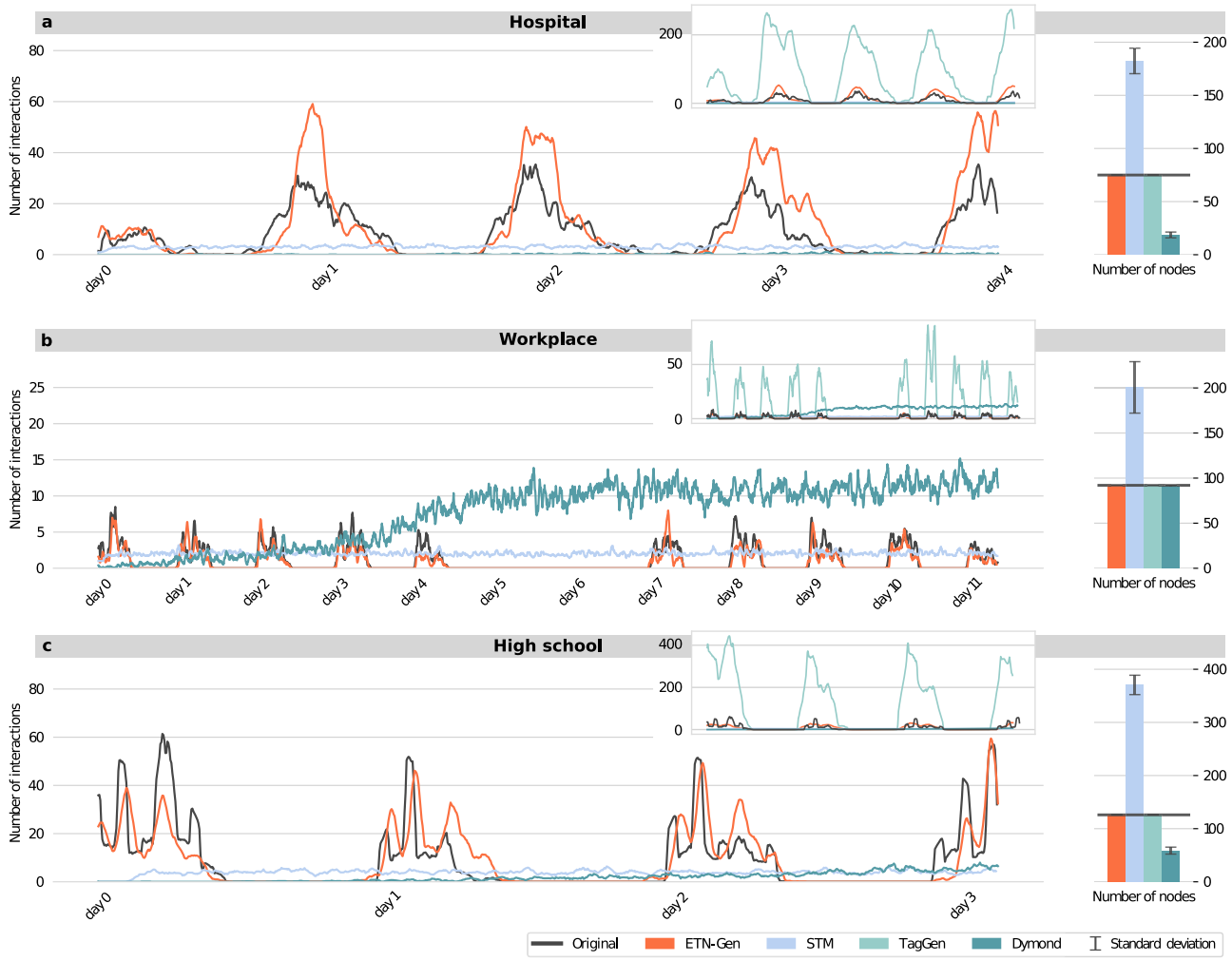

Fig. 2 Number of interactions in time and number of nodes in the original and surrogate networks. Each color represents a different generation algorithm, while the original graph is depicted in black. The insets depict the same curves with a different *y*-scale for visibility (the results obtained for *TagGen* are only reported there). The bar plots represent the number of nodes for each generation algorithm with standard deviations in gray. Panels **a**, **b**, and **c** correspond to Hospital, Workplace, and High school networks, respectively.

as the original graph, which allows us to statistically preserve the overall average number of interactions of the original graph. Moreover, we manually input periodicity via different local probabilistic models for different times and days of the week. We highlight, however, that while using only a single local probabilistic model would remove our ability to model periodic changes in graph over time, we would still be able to model the average number of interactions, as these are automatically reproduced by the rest of the algorithm. A detailed analysis is reported in the Supplementary Note 6.

From the comparison with the other methods we conclude that *ETN-gen* is the only method able to preserve the number of nodes, the order of magnitude of the amount of interactions and the periodicity of the original network. The curves for the original network and *ETN-gen* are also reported in Supplementary Figure 11 for larger datasets to which the other methods cannot be applied due to computational constraints (see Supplementary Note 7).

Topological similarity evaluation. Having studied the temporal development, we now turn to structural similarity between the surrogate data and the original networks. We consider seventeen metrics for structural similarity, divided between those that depend on time, namely: number of connected components[44], density[33], number of interacting individuals[11], new conversations[11], hour S-metric[45], hour modularity[46,47], duration of contacts[11], closeness[44], hour betweenness (weighted and unweighted)[44], hour clustering[48,49], hour assortativity[50], hour average shortest path length[1]; and those that are measured on the aggregated network (i.e. collapsing all the temporal layers in one weighted network), which are: closeness[44], betweenness (weighted and unweighted)[44], and edge strength[11]. All the measures are collected as distributions, the temporal ones measured on each singular temporal layer, and the aggregated ones as distributions over the edges or the nodes. Among those measured on singular temporal layers, the ones denoted with Hour are measured after having increased the aggregation time of temporal layers to 1 hour. This was necessary to increase density and obtain measures otherwise not meaningful. See Supplementary Note 8 for the definition of all measures.

To compare distributions we rely, inspired by Zeno et al.[33], on the Kolmogorov–Smirnov distance[51] to contrast generated and original graphs. In the Supplementary Note 9, we also consider alternative distance measures, namely the Jensen–Shannon divergence[52], the Kullback–Leibler divergence[53], and the Earth mover's distance[54], obtaining similar results (see Supplementary Note 9). Distances between distributions are reported in Figs. 3 and 4, where we compare graphs obtained with *ETN-gen* with

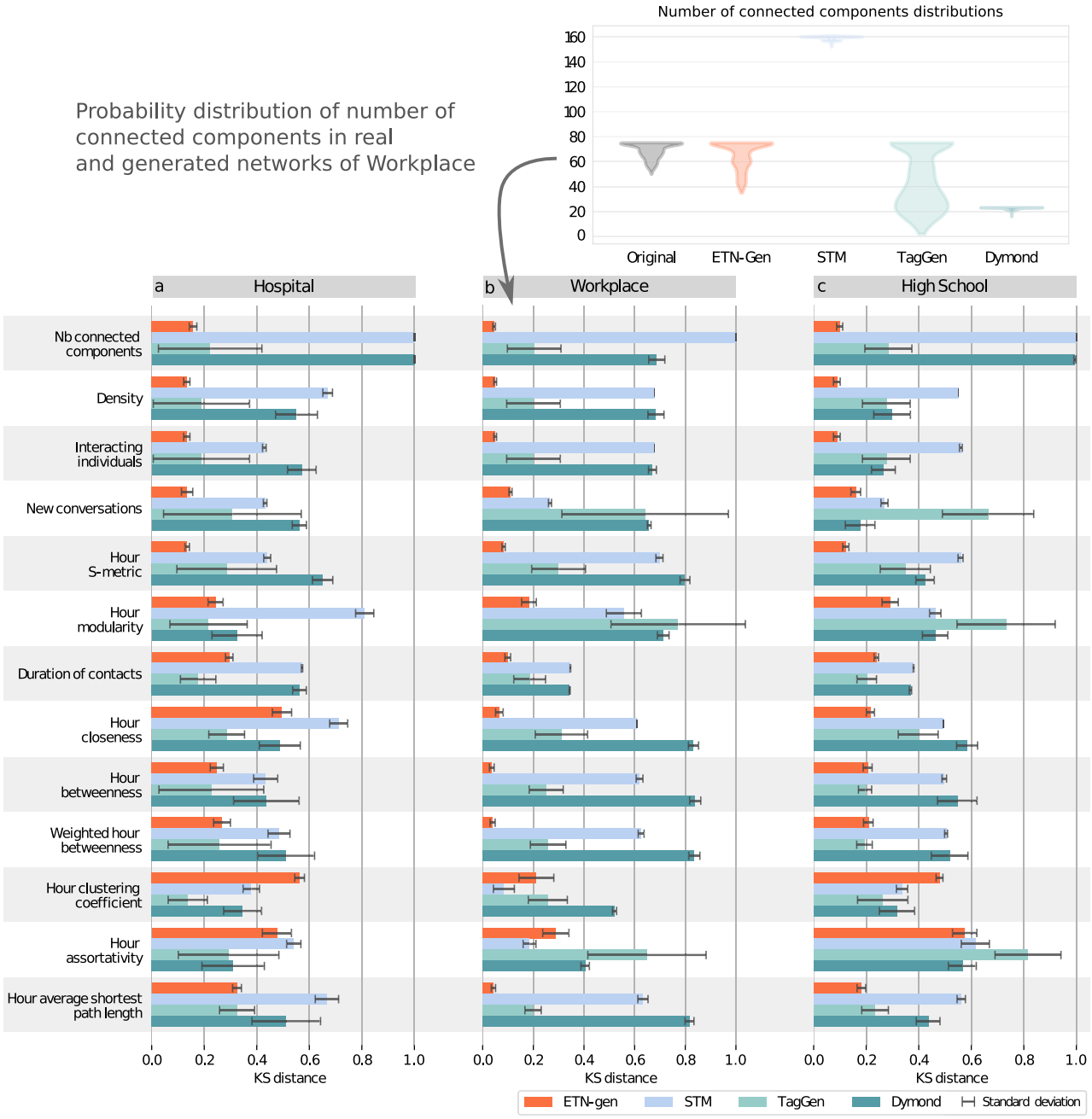

Fig. 3 Topological similarity according to time-dependent measures. Similarity of the original network with those generated by Egocentric Temporal Neighborhood Generator (*ETN-gen*), Structural Temporal Modeling (*STM*), *TagGen* and DYnamic MOtif-NoDes (*Dymond*). Each bar reports the Kolmogorov–Smirnov distance between the two distributions (original and generated) for a specific structural metric. The shorter is a bar the more similar are the distributions. Standard deviations are obtained over 10 stochastic realizations of each network. In the top inset we report the distributions of the number of connected components in real and in one instance of generated networks for the Workplace dataset. Panels **a**, **b**, and **c** correspond to Hospital, Workplace, and High school networks, respectively.

those from the three alternative approaches. The networks generated with *ETN-gen* (orange bars) show a high similarity to the original networks for many of the measures and also a high stability (small errorbars). The measures for which *ETN-gen* performs best are those that, together with the number of interactions (see Fig. 2), are preserved by construction: the density and the number of interacting individuals in time. Here, the similarity originates from the neighborhood probability distributions, which ensure that from a statistical viewpoint, the surrogate network has the same number of interactions and the same number of individuals involved in an interaction. The same holds for the number of times that a new link appears, as these statistics are also stored in the neighborhood probability distributions. Another characteristic that is well captured by the egocentric temporal neighborhoods is the hub-like structure that we can find in each static layer, which is measured by the S-metric[45].

Going beyond these trivial consequences of the mechanics of the generating mechanisms, the method does well at preserving the number of connected components. Indeed, the inset shows how the *ETN-gen* networks exhibit a distribution of the number of connected components that is similar to the original one, while

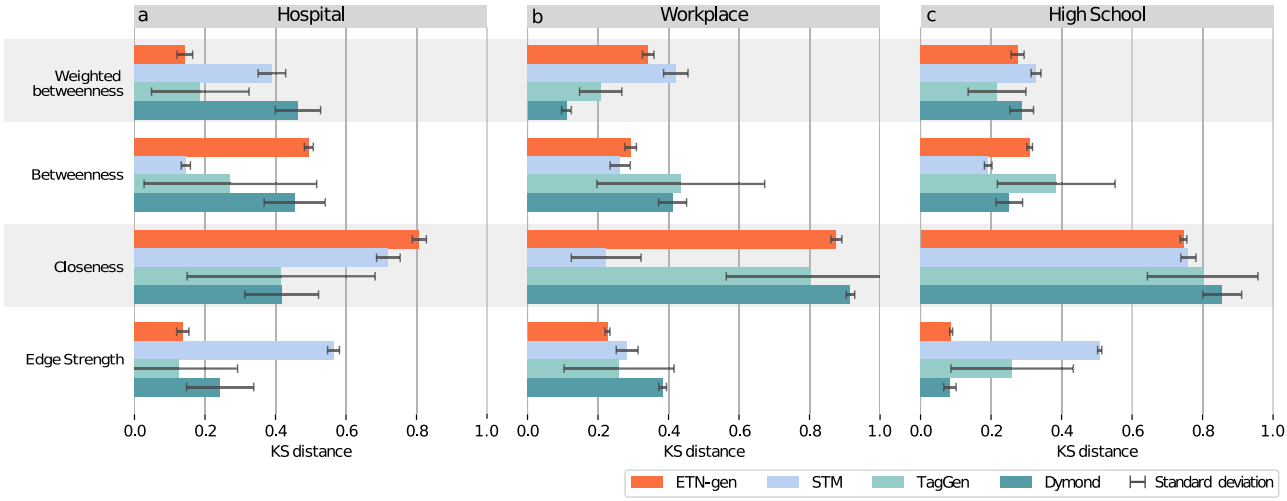

Fig. 4 Topological similarity according to time-aggregated measures. Similarity of the original network with those generated by Egocentric Temporal Neighborhood Generator (*ETN-gen*), Structural Temporal Modeling (*STM*), *TagGen* and DYnamic MOtif-NoDes (*Dymond*). Each bar reports the Kolmogorov–Smirnov distance between the two distributions (original and generated) for a specific structural metric. The shorter is a bar the more similar are the distributions. Standard deviations are obtained over 10 stochastic realizations of each network. Panels **a**, **b**, and **c** correspond to Hospital, Workplace, and High school networks, respectively.

TagGen shows a rather larger distribution difference and the other methods generate substantially fewer (*Dymond*) or more (*STM*) connected components. This is a consequence of the fine-grained temporal information that *ETN-gen* uses to generate the networks. For the same reason, the hour modularity of *ETN-gen* networks is always better or comparable with that of the other generated networks. The distribution of durations is instead not very well reproduced, but it is not bad with respect to the other methods. In fact, considering a k-steps memory allows interactions to have a continuity in time, differently from the case of independent layers. We also test three different centrality measures. Since centrality is a quantity that characterizes each node at each time, we focus first on the temporal distributions, by reporting for each temporal slot lasting one hour the nodes average (see Fig. 3). Then we consider the spatial distribution, reporting the centrality of each node computed on the time aggregated network (see Fig. 4). We observe that when we consider the temporal distribution we obtain a higher similarity to the original networks, confirming the insight that *ETN-gen* is more valid in reproducing fine-grained time features, while it results more limited in reproducing the global spatial organization of the networks.

Another property is the distribution of edge strengths in the projected graph (see Fig. 4). Edge strength is simply the number of times that each edge has appeared over the duration of the graph. Here, we would not necessarily expect *ETN-gen* to do well as the method will tend to create networks with quite homogeneous distributions of strength. This is because it can only rely on a memory of order k for edge repetitions, and does not have a long-term memory. Hence all the heterogeneous behaviors that we can find for instance in social datasets, where individuals tend to establish relationships with specific nodes and have repeated (but not necessarily consecutive) interactions with them, are not preserved by *ETN-gen*. Nevertheless, we find that for the considered datasets *ETN-gen* remains competitive with the other methods.

If edge strength is partially affected by the absence of long memory, the most important limitations of the egocentric perspective are highlighted by clustering, degree assortativity and average shortest path length, which are related to second-order interactions (see Fig. 3). This is the cost we pay for having a computationally efficient model applicable to arbitrary networks. Notice that while this is a problem in theory, it seems not to affect the workplace dataset, which is a substantially sparser network with low clustering and short paths.

In general from the topological analysis we observe that the features that are more preserved are the time-dependent ones (at least those that do not depend on second-order interactions), while the method is more limited in reproducing the time-aggregated measures. This is valid for both local and global features. An additional analysis on meso-scale structures is reported in Supplementary Note 10. We observe that small static motifs are well preserved by *ETN-gen* but not the network communities (if present in the original graph). This is a common limitation involving the other generation methods, too.

Dynamical similarity evaluation. Having tested our method from the structural point of view, we now test the usefulness of the surrogate networks in terms of dynamical processes unfolding upon them. We study two dynamical models: random walk and a spreading model.

Random walk. We simulate a temporal random walk[11,55] on the original and generated networks. We use the standard definition of random walk extended to temporal networks: a random walker starts from a randomly chosen node at generic time t and chooses uniformly at random one of its neighbors, moving there. Then the second step will take place on the following layer of the temporal network, so the walker will randomly choose between its neighbors at time $t+1$, and so on, assuming that at each time corresponds one and only one jump. We compute two metrics: coverage and mean first passage time (MFPT), and compare distributions over different realizations between the input and the generated temporal network using again the Kolmogorov–Smirnov distance (see Supplementary Note 9 for the definition of the metrics). We consider three different starting points: $t=0$, $t=len(G)/2$ (in the middle point of the temporal extension of the graph), and the time corresponding to the first peak of connections (when the number of connections reaches a maximum).

In Fig. 5 we report the Kolmogorov–Smirnov distance for coverage and MFPT. The horizontal dashed line shows the

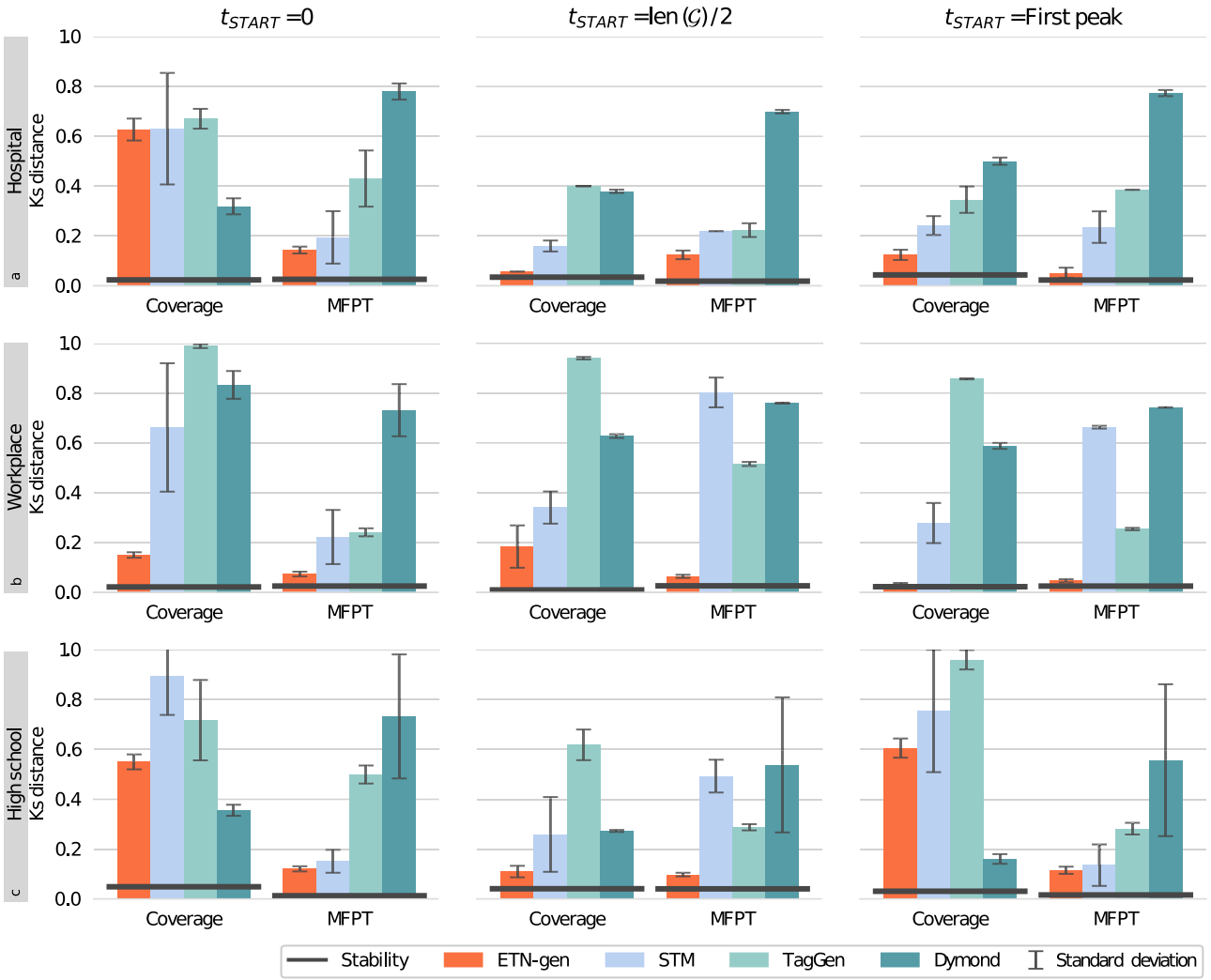

Fig. 5 Dynamic similarity: random walk. Kolmogorov–Smirnov distance between original and generated distributions of coverage and mean first passage time (MFPT) in the random walk model in each generated network for three different starting points: time 0, T/2 and on the first peak. Our method is represented in orange, while the solid black line shows the stability (i.e. the same simulation on the original network). Panels **a**, **b**, and **c** correspond to Hospital, Workplace, and High school networks, respectively.

stability of each measure on the original network. The black line is obtained comparing different performances (average over 1000 simulations of random walk for coverage, and 5 times each couple of nodes for mean first passage time) by means of the Kolmogorov–Smirnov distance. It is worth noting that the stability is different from zero, due to inherent variations within the dynamic process. We observe that the dynamics on the *ETN-gen* networks are similar to the ones on the original networks in terms of mean first passage time, while in terms of coverage performance they depend on the datasets and the starting point but they always results competitive with the ones obtained with the alternative methods.

In Supplementary Note 11 we also report the evolution in time of the number of newly visited nodes. In general we can say that the random walk process on the *ETN-gen*'s surrogate networks is quite similar to the random walk on the original graph.

Spreading model. We simulate a Susceptible-Infectious-Recovered (SIR) model[56], with three possible values for the probability of disease transmission ($\lambda \in \{0.25, 0.13, 0.01\}$), and the recovery rate fixed at $\mu = 0.055$. In each simulation the infection starts at time t_{START} by assigning to one random node (selected among the connected ones at that time) the status of infected. This initial node will infect its neighbors with probability λ and recover with probability μ. In the next time step we consider the following temporal layer of the network and again the infected nodes can infect their new neighbors or they can recover. We repeat the procedure until the end of the temporal layers or until all the infected nodes have recovered. Again, we consider the three different starting points described above also for the dynamics. We compute the reproduction value R_0. Each experiment was repeated 100 times and the distribution of R_0 obtained on the original network is, again, compared with those obtained on synthetic networks by means of the Kolmogorov–Smirnov distance. Results are shown in Fig. 6, where again a horizontal black line shows the stability of each measure on the original network (computed averaging over 100 simulations). The evolution in time of the number of infected nodes is reported in Supplementary Note 11. The computation of R0 is affected by a large variability, both on *ETN-gen* networks and on the other generated networks. This is probably due to the limitations of all these methods, up to now incapable to reproduce several meso-scale properties, like modularity, clustering, and temporal correlations (see Supplementary Note 10 for an exploration of these features). Future works will be devoted to fill these gaps.

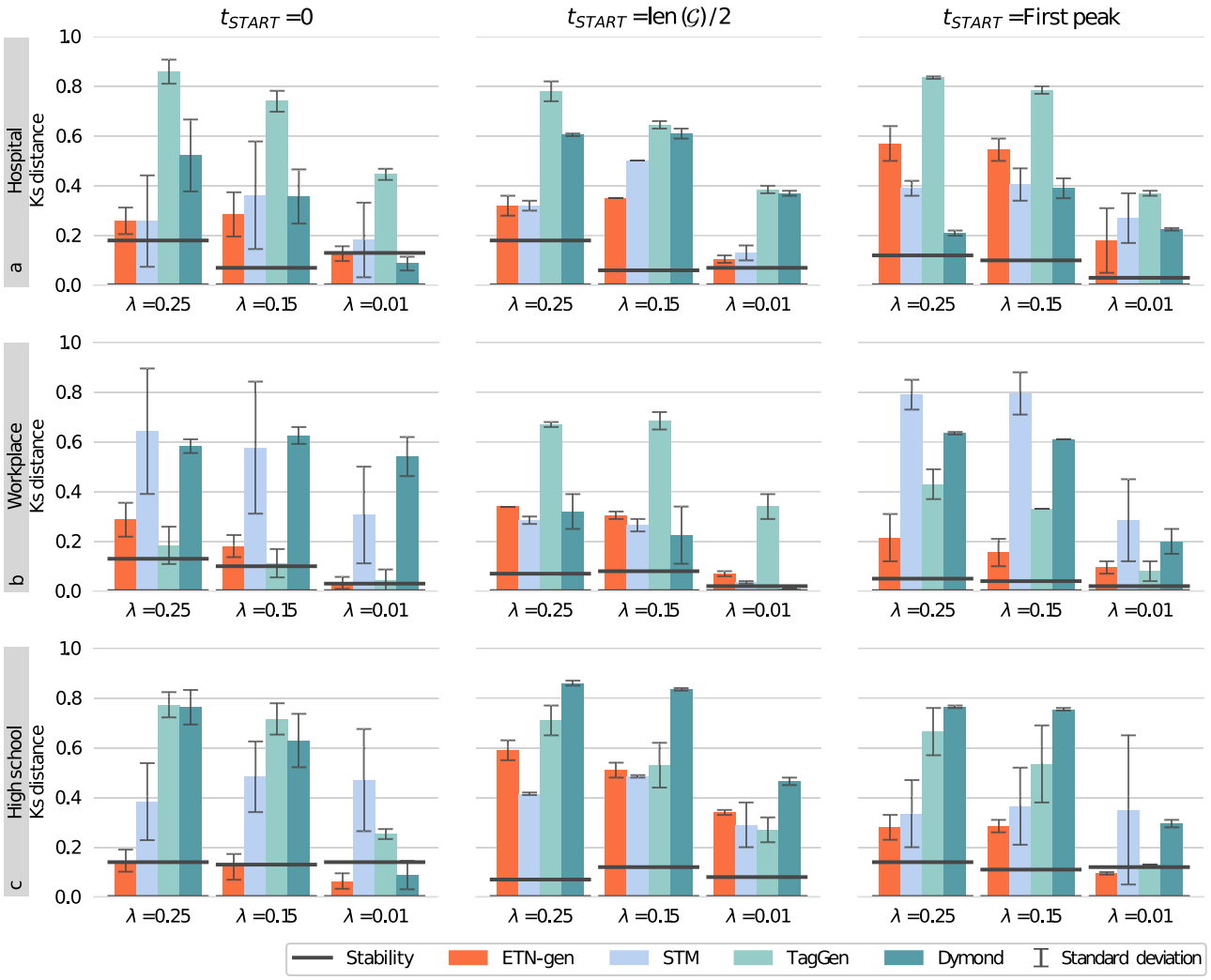

Fig. 6 Dynamic similarity: spreading model. Kolmogorov–Smirnov distance between original and generated distributions of R_0 values on a Susceptible-Infected-Recovered (SIR) model simulation in each generated network for three different starting points: time 0, T/2 and on the first peak. Our method is represented in orange, while the solid black line shows the stability (i.e. the same simulation on the original network). Panels **a**, **b**, and **c** correspond to Hospital, Workplace, and High school networks, respectively.

Dataset expansion and extension. In the previous sections we have argued that *ETN-gen* creates realistic surrogate temporal networks that mimic many aspects of real social dynamics (both in terms of structure and in reproducing dynamical systems).

Now we ask the question: How can this tool be useful in practice? A relevant application is represented by the possibility of enlarging a given temporal dataset, both in time and in size. It is indeed common that a specific analysis, in order to yield reliable results, requires a larger population or a longer time than those characterizing collected real data. In those cases we deal with the long-standing problem of data augmentation, for which we now argue that *ETN-gen* represents a promising solution. In the following we show how our method can be used for augmenting a temporal dataset, by adding temporal layers (temporal extension), but also by increasing the size of the network in terms of number of nodes (size expansion).

Temporal extension. The procedure, as explained in the previous sections, implies calculating the neighborhood probability distributions, which somehow summarizes the interaction patterns in the original graph. Each layer of the surrogate networks is built extracting possible interactions for each node from these distributions, a process that only depends on the last k layers. The temporal extension of a dataset is therefore straightforward: the procedure of temporal layer addition can be repeated possibly an infinite number of times, and we stop when the desired number of time steps is reached. At the top of Fig. 7 we show an example of temporal extension of the workplace network. We have selected this dataset to highlight the ability of *ETN-gen* to differentiate between week days and weekends. To evaluate the quality of the extension, we assume to only know the first week of the original two-week dataset (from the beginning to the vertical line) and from this we estimate the neighborhood probability distributions. We then use it to generate an ensemble of 10 surrogate networks with a length of two weeks. The mean and standard deviation of the number of interactions in the generated graph are reported in orange. The number of interactions of the real graph are reported in black dashed curves for the first week, and in black solid curves for the following week. In other words, the method is trained on the first week of the real dataset, several two-week networks are generated, and they are eventually tested comparing them with the two-weeks-long real dataset. Results show how the generated networks accurately recreate the original behavior beyond the timespan that was used to estimate the local probability distributions.

Size expansion. Here we explore the fidelity of surrogate networks with an increased number of nodes. As discussed above, it is

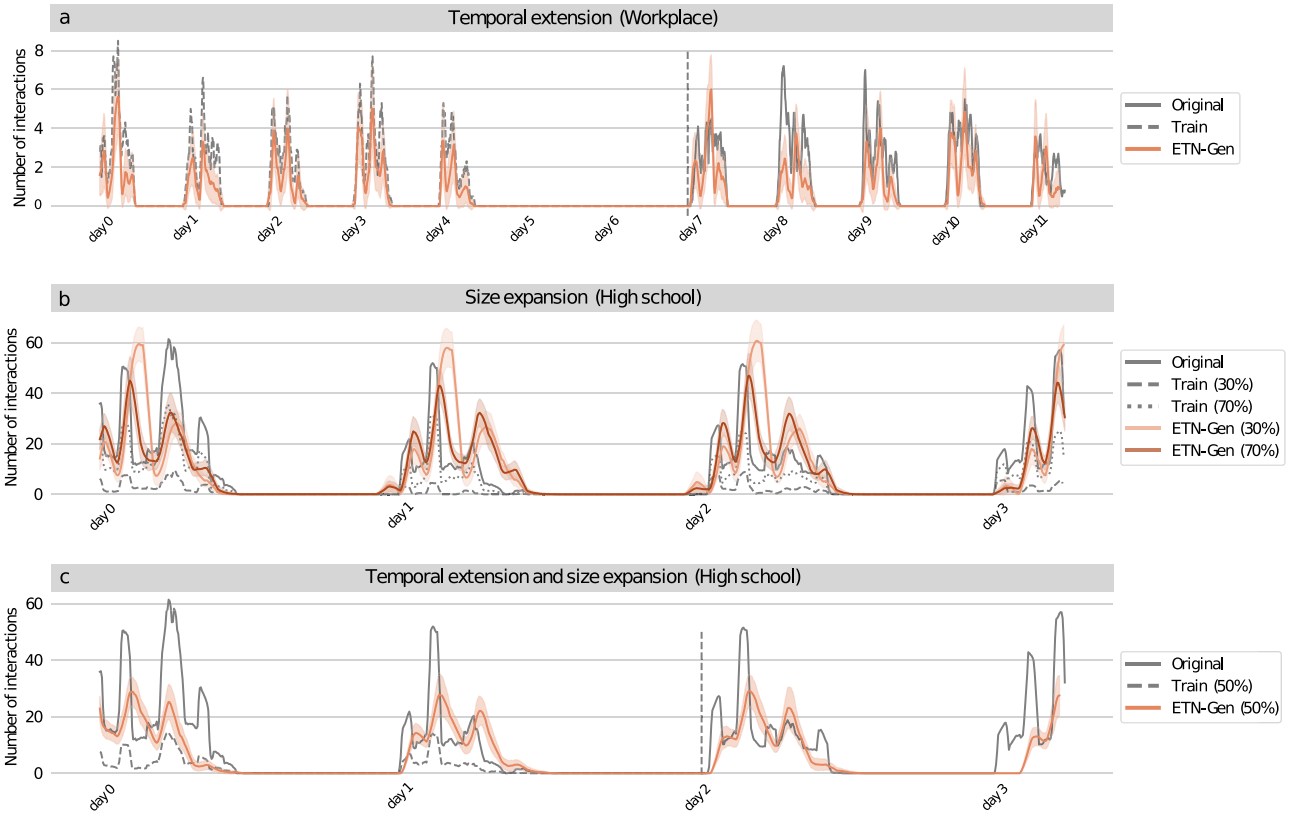

Fig. 7 Temporal extension and node expansion. Panel **a** displays the temporal extension within the Workplace network. Panel **b** illustrates the size expansion in the High school network. Panel **c** shows both the temporal extension and size expansion in the High school network. The mean and standard deviation of our method are shown in orange (and brown). Black dashed (and dotted) lines show the original data used to train our model, while black solid lines show the original data used to evaluate the quality of the generated network. In experiments involving temporal expansion, a vertical bar separates the temporal range used to collect training data from the one where expansion is performed.

possible to increase the size beyond that of the original network within the *ETN-gen* framework because the number of nodes is simply a parameter to set for the method. That said, however, the concept of size expansion requires more attention than time extension. Because, as we change the number of nodes in a network we should also consider how the density of the graph and the mean degree should change accordingly.

In the following we describe an experiment of data augmentation, assuming that we only have access to incomplete data. Incomplete data are obtained by randomly removing part of the nodes from the original network. We use the high school dataset which, with its 126 nodes, is the largest among our datasets, and we consider two reduced versions, with 30% and 70% of the nodes respectively. When removing part of the nodes from a network, we naturally remove also part of the links (all those which were before connecting the eliminated nodes to the remaining ones), we hence reduce the mean degree. We should consider that an incomplete dataset has in general a reduced mean degree with respect to the real-world network, and that when we try to reconstruct the original network via data augmentation we should increase the mean degree too. See Methods for a quantification of the needed increase.

Anyway, once the desired connectivity has been chosen, *ETN-gen* allows us to generate a surrogate network with the desired number of nodes and the desired degree, while maintaining the pattern of egocentric interactions of the original dataset.

The results of the experiment on the high school dataset are shown in panel a of Fig. 7. For each of the two reduced temporal networks we generate a temporal network with 126 nodes to try to reconstruct the original graph. We generate the initial snapshot using the configuration model based on the degree distribution of the first snapshot of the original (not reduced) graph. Then we build local probability distributions only using information from the reduced networks and use these local probability distributions to generate surrogate expanded networks from them. The expanded networks have the same number of nodes of the original one (126), enabling direct comparison. The expedient that we use to augment the mean degree from the reduced seed graph is to increase the parameter α of the generation process, which is the probability to confirm the unidirectional directed links in each provisional layer (set to 1/2 by default). See Methods for the details on how to compute the correct value of α given the original number of links and the desired density of the generated graph.

Panel b of Fig. 7 the black solid curve represents the number of interactions in the original network, the black dashed curve those in the train network with 30% of the nodes and the black dotted curve those in the one with 70% of the nodes. The corresponding values for the generated networks with their standard deviations are reported in orange and brown respectively. Again, we observe the ability of our method to correctly replicate the pattern of interaction in the original network, even if fed with a small percentage of nodes from the original graph as seed.

Temporal extension and size expansion. We can also combine the two techniques above to simultaneously increase the number of nodes and the temporal snapshots. The results are shown in panel c of Fig. 7 for the high school network, where the synthetic graph has been obtained by only using 50% of the nodes and the first two days of the original dataset (from the beginning to the vertical line), see the

black dashed curve. Also in this case, our method can expand an input graph in both temporal and node size dimensions.

Discussion

In this manuscript we have proposed a model to generate surrogate temporal networks, i.e. synthetic networks that realistically capture many properties of real-world datasets, only making use of the information contained in egocentric temporal neighborhoods. Specifically, we generate temporal networks which accurately reproduce structural characteristics like density, number of interacting individuals, number of connected components, and the possible presence of hubs. We observe that in both topological and dynamical tests, the networks generated by this model are generally closer to the original graph than those generated by different literature models.

Moreover, this approach is able to generate temporal networks that have different sizes than the original one. This property can be used to increase the number of nodes and extend the network in time, providing a powerful tool for data augmentation. These results suggest that egocentric temporal neighborhoods, that we use as building blocks, contain fundamental information about the real networks they are extracted from.

By using *ETN-gen* surrogate networks it is possible to overcome privacy issues, too. We did not explicitly prove that such surrogate networks are impossible to de-anonymize, but we are rather confident in the privacy-preserving properties of the method. In fact, the interactions of one node in the surrogate are designed based on the probability distribution of ETN prolongation, that is in its turn constructed based on the interactions of all the nodes in the original graph (remembering that the identity of nodes are not stored). Therefore there is not a match node-to-node between surrogate and original graph. More precisely, the set of interactions of one node in the original graph is distributed among multiple nodes in the surrogate graph. We hence find it unlikely that real nodes can be reconstructed and identified observing the surrogate network. However, this will be matter of future investigations.

The other side of the coin is that this simplicity does not capture certain topological features. This is the main limitation of the model. For instance, disregarding second-order interactions translates to a reduced ability to preserving clustering, degree correlations and average shortest path length. This is the price to pay to achieve scalability, sidestepping the graph isomorphism problem in mining egocentric temporal neighborhoods. Just getting rid of this simplifying assumption would imply a substantial blow-up in computational complexity (as suggested by the runtime comparisons with alternative approaches reported in Supplementary Table S1) and, as a consequence, a significant reduction in the timespan of the temporal neighborhood that could be dealt with. Trading second order interactions for longer temporal neighborhoods allows us to reproduce most of the relevant features of temporal networks while maintaining computational efficiency. Nevertheless, further research is needed to explore alternative trade-offs in the expressivity-efficiency scale. Another limitation of the proposed approach is the absence of long-term memory, which implies that the model cannot capture long-term patterns of interaction (like e.g. daily or weekly recurrences). These features are instead well captured by more theoretical models of network generation that include aging[57,58], edge reinforcement[59,60], or in general some mechanism for memory such that contact duration and inter-event times are heterogeneous and depend on the past interactions[61,62]. Memory could also be used to generate a synthetic temporal network that is organized in communities[63,64]. This is a characteristic often occurring in social networks (particularly evident in schools), and it cannot be captured by small local subnetworks like egocentric temporal neighborhoods. However, long-term memory appears in literature only in theoretical models for temporal network generation, for which the goal is to obtain realistic networks by recovering some particular characteristics of the observed dynamics in real networks, but usually do not aim at reconstructing specific real networks or environments. Indeed, the alternative methods we evaluated in this manuscript also fail to account for long-term memory. A model which instead is built to obtain surrogate networks with an alternative approach is the one proposed by Presigny et al.[18]. This model does not generate a new network from scratch, it instead individuates a backbone of a real temporal network, defined as the global subnetwork composed of the most significant edges, and then reconstructs the missing links. This is based on a conceptually different idea, assuming that the important information concerns the global structure of the network, while the method that we are proposing focuses on how nodes behave given their interactions in last time steps. This is indeed evident from Figs. 3 and 4: if *ETN-gen* is highly effective in reproducing the evolution in time of singular node neighborhoods, it fails in reproducing global network features that are not reproducible by only using ego-node information. By recalling two different long-standing traditions in network science, a sociocentric versus an ego-centric perspective[49], we can assert that if the first one is covered, for what concerns surrogate temporal networks, by the model of Presigny et al.[18], our model places itself in the remaining gap, filling the unexplored case of the egocentric perspective.

The insertion of memory or second order mechanisms, implying the possibility to reconstruct an organization in groups of nodes, and also to make the set of nodes change in time, inserting new nodes or excluding old ones, are demanded to future work, aiming at improving the current method for a further advance in realistic reconstruction.

Methods

Data description and processing. The three temporal networks studied in the main body of this work represent face-to-face human interactions collected by the SocioPatterns project:

- Hospital[37]. The dataset has been collected in the geriatric ward of a university hospital[65] in Lyon, France, over four days in December 2010. It contains interactions among medical doctors, paramedical staff, administrative staff and patients. Number of edges: 1139, number of nodes: 75.
- Workplace[38]. The dataset has been collected in 2013 at the Institut National de Veille Sanitaire, a health research institute near Paris, over two weeks. It contains interactions among individuals from five departments. Number of edges: 755, number of nodes: 92.
- High school[39]. The dataset has been collected in 2011 in Lycée Thiers, Marseilles, France, over four days (Tuesday to Friday). It contains interactions among 118 students and 8 teachers in three different high school classes. Number of edges: 1709, number of nodes: 126.

As stated by the researchers involved in the aforementioned studies, each study participant and staff member was asked to sign an informed consent and each study received the approval by the French national body responsible for ethics and privacy, namely the "Commission Nationale de l'Informatique et des Libertés" (CNIL, http://www.cnil.fr). More details can be found in the publications describing the studies and the collected data[37–39].

Kolmogorov–Smirnov distance. The (two-sample) Kolmogorov–Smirnov test[51] is a non-parametric test used to test

how likely it is that two sets of samples come from the same (unknown) distribution. The test uses the following statistic:

$$D_{KS} = \max_x |F_1(x) - F_2(x)|$$

Where $F_1(x)$ and $F_2(x)$ are the empirical cumulative distributions of the two sets.While originally conceived for hypothesis testing, the KS statistic has often been used to measure the distance between empirical cumulative distributions[33,66–70]. We follow this common practise in this manuscript.

Neighborhood generation process: parameters. The gap between two consecutive temporal snapshots has been set to 5 minutes for face-to-face interaction networks and 10 minutes for SMS and phone call networks (in Supplementary Note 2 and 3). The time horizon k defining the egocentric temporal neighborhood has been set to $k = 2$ in all experiments, which is the minimal horizon that preserves some temporal correlation. In Supplementary Note 12 we motivate our decision in using $k = 2$ and we also show the results for $k = 3$ in Supplementary Note 13. Local probability models have a granularity of 1 hour and a periodicity of 1 day (i.e. between 8 and 9 am in each day we use the same probability model, and the same holds for all 1 hour slots in the day), for all networks but the ones including weekends, namely Workplace and High school 2, for which the periodicity is set to 1 week.

Space and time complexity. The time complexity required by our method is $\mathcal{O}(n \cdot m)$, where n is the number of nodes in the temporal graph and m is the number of timestamps. The space complexity is constant with respect to both network size and number of timestamps. See Supplementary Note 14.

Size expansion: preserving interaction density. The seed graphs for the size expansion experiment are generated by artificially reducing the original dataset (so that the original graph can be used as ground-truth). In this reduction process, whenever a node is dropped all its connections are dropped too. As a consequence, the resulting seed graph has a reduced mean degree with respect to the original one, and the expanded graph generated from it would inherit this reduced mean degree. This problem can be avoided by adjusting the α parameter of the generation process (the probability to confirm the unidirectional links in each provisional layer, set to 1/2 by default). In particular, we would need to set $\alpha = 1 - \frac{1}{2}\frac{\hat{L}}{L}$, where $\hat{L}$ is the average number of links in the seed graph and L the desired number of links in the generated graph. However, L is unknown and needs to be estimated. Something that we know, and that we want in this case to preserve, is the density, defined as $d = \frac{\hat{L}}{\hat{N}\cdot(\hat{N}-1)/2}$ i.e. the fraction between the number of links in the seed graph and all possible links ($\hat{N}$ is the number of nodes in the seed graph). If we assume a linear growth with respect to the number of all possible edges in the network, we also have: $d = \frac{L}{N\cdot(N-1)/2}$, with N as the number of nodes of the generated graph (that we can choose). Combining these two equations we obtain an estimate for L, from which we obtain: $\alpha = 1 - \frac{N\cdot(\hat{N}-1)}{\hat{N}\cdot(N-1)} \cdot \frac{1}{2}$. Hence, when we consider a seed with only 30% of the nodes of the high school dataset (so $N = 126$ and $\hat{N} = 38$) we should use $\alpha = 0.96$ to reproduce the same density. While if we start with 50% and 70% of the nodes (i.e. $\hat{N} = 63$ and $\hat{N} = 88$) in the seed we should use respectively $\alpha = 0.88$ and 0.76.

Alternatives approaches for generating networks. *Dymond*[33] builds a temporal network considering (i) the dynamics of temporal motifs in the graph and (ii) the roles nodes play in motifs (e.g. in a wedge – two links connecting three nodes – one node plays the hub, while the remaining two act as spokes). The method has no parameters to be set. *STM*[34] extracts counts for a predefined library of (non-egocentric) temporal motifs from the original network, and turns them into generation probabilities from which to create the temporal network. In particular, we use the parameterized version of *STM* and we set the parameter $\alpha = 0.6$ as recommended by[34]. *TagGen*[35] is a neural-network based approach that extracts temporal random walks from the original graph and feeds them to an assembling module for generating temporal networks. *TagGen* has been trained with the parameters used in the original paper, namely 30 epochs with a batch size of 64 and stochastic gradient descent with a learning rate of 0.001.

Data availability

The data used to support this study are publicly available at the following links. • The SocioPatterns data[37–39] at http://www.sociopatterns.org • The CNS data[17] at https://doi.org/10.6084/m9.figshare.7267433 • The Friends and Family data[16] at http://realitycommons.media.mit.edu/friendsdataset.html.

Code availability

The codes used for the generation of temporal network are publicly available at the following links. • *ETN-gen*: https://github.com/AntonioLonga/ETNgen • *STM*: https://github.com/temporal-graphs/STM • *TagGen*: https://github.com/davidchouzdw/TagGen • *Dymond*: https://github.com/zeno129/DYMOND.

Received: 14 January 2023; Accepted: 27 December 2023;
Published online: 09 January 2024

References

1. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
2. Han, J.-D. J. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
3. Eagle, N. & Pentland, A. Reality mining: sensing complex social systems. *Personal. Ubiquitous Comput.* **10**, 255–268 (2008).
4. Chechik, G. et al. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat. Biotechnol.* **26**, 1251–1259 (2008).
5. Balcan, D. et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci.* **106**, 21484–21489 (2009).
6. Cattuto, C. et al. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS One* **5**, e11596 (2010).
7. Starnini, M. & Pastor-Satorras, R. Topological properties of a time-integrated activity-driven network. *Phys. Rev. E* **87**, 062807 (2013).
8. Corsi, F., Lillo, F., Pirino, D. & Trapin, L. Measuring the propagation of financial distress with granger-causality tail risk networks. *J. Financial Stab.* **38**, 18–36 (2018).
9. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320 (2019).
10. Ciaperoni, M. et al. Relevance of temporal cores for epidemic spread in temporal networks. *Sci. Rep.* **10**, 12529 (2020).
11. Starnini, M., Baronchelli, A., Barrat, A. & Pastor-Satorras, R. Random walks on temporal networks. *Phys. Rev. E* **85**, 056115 (2012).
12. Rocha, L. E., Masuda, N. & Holme, P. Sampling of temporal networks: Methods and biases. *Phys. Rev. E* **96**, 052302 (2017).
13. Cencetti, G. et al. Digital proximity tracing on empirical contact networks for pandemic control. *Nat. Commun.* **12**, 1–12 (2021).
14. Isella, L. et al. Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One* **6**, e17144 (2011).
15. Stehlé, J. et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC Med.* **9**, 1–15 (2011).
16. Aharony, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput.* **7**, 643–659 (2011).
17. Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Lehmann, S. Interaction data from the copenhagen networks study. *Sci. Data* **6**, 1–10 (2019).

18. Presigny, C., Holme, P. & Barrat, A. Building surrogate temporal network data from observed backbones. *Phys. Rev. E* **103**, 052304 (2021).
19. Mollgaard, A., Lehmann, S. & Mathiesen, J. Correlations between human mobility and social interaction reveal general activity patterns. *PLoS One* **12**, e0188973 (2017).
20. Kikas, R., Dumas, M. & Karsai, M. Bursty egocentric network evolution in skype. *Soc. Netw. Anal. Min.* **3**, 1393–1401 (2013).
21. Cretu, A.-M. et al. Interaction data are identifiable even across long periods of time. *Nat. Commun.* **13**, 1–11 (2022).
22. Bois, F. Y. & Gayraud, G. Probabilistic generation of random networks taking into account information on motifs occurrence. *J. Computational Biol.* **22**, 25–36 (2015).
23. Coscia, M. & Szell, M. Multiplex graph association rules for link prediction. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media* (ICWSM, 2021).
24. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
25. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629 (2000).
26. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633 (2000).
27. Bianconi, G. & Barabási, A.-L. Competition and multiscaling in evolving networks. *EPL Europhys. Lett.* **54**, 436 (2001).
28. D'souza, R. M., Borgs, C., Chayes, J. T., Berger, N. & Kleinberg, R. D. Emergence of tempered preferential attachment from optimization. *Proc. Natl Acad. Sci.* **104**, 6112–6117 (2007).
29. Papadopoulos, F., Kitsak, M., Serrano, M., Boguná, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537–540 (2012).
30. Perra, N., Gonçalves, B., Pastor-Satorras, R. & Vespignani, A. Activity driven modeling of time varying networks. *Sci. Rep.* **2**, 1–7 (2012).
31. Gauvin, L. et al. Randomized reference models for temporal networks. *SIAM Review* **64**, 1–7 https://doi.org/10.1137/19M1242252 (2022).
32. Berlingerio, M., Bonchi, F., Bringmann, B. & Gionis, A. Mining graph evolution rules. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 115–130 (Springer, 2009).
33. Zeno, G., La Fond, T. & Neville, J. Dymond: Dynamic motif-nodes network generative model. In *Proceedings of the Web Conference 2021*, 718–729 (ACM, 2021).
34. Purohit, S., Holder, L. B. & Chin, G. Temporal graph generation based on a distribution of temporal motifs. In *Proceedings of the 14th International Workshop on Mining and Learning with Graphs*, vol. 7 (ACM, 2018).
35. Zhou, D., Zheng, L., Han, J. & He, J. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 401–411 (ACM, 2020).
36. Grohe, M. & Schweitzer, P. The graph isomorphism problem. *Commun. ACM* **63**, 128–134 (2020).
37. Vanhems, P. et al. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One* **8**, e73970 (2013).
38. Génois, M. et al. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Netw. Sci.* **3**, 326–347 (2015).
39. Fournet, J. & Barrat, A. Contact patterns among high school students. *PLoS One* **9**, e107878 (2014).
40. Longa, A., Cencetti, G., Lepri, B. & Passerini, A. An efficient procedure for mining egocentric temporal motifs. *Data Min. Knowl Discov.* **36**, 355–378 (2022).
41. Molloy, M. & Reed, B. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* **6**, 161–180 (1995).
42. Newman, M. E., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
43. Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling tabular data using conditional GAN. *Adv. Neural Inf. Proces. Syst.* **32**, 7335–7345 (2019).
44. Newman, M. *Networks* (Oxford university press, 2018).
45. Li, L., Alderson, D., Doyle, J. C. & Willinger, W. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.* **2**, 431–523 (2005).
46. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
47. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).
48. Luce, R. D. & Perry, A. D. A method of matrix analysis of group structure. *Psychometrika* **14**, 95–116 (1949).
49. Wasserman, S. & Faust, K. *Social network analysis: Methods and applications* (Cambridge University Press, 1994).
50. Newman, M. E. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
51. Massey Jr, F. J. The kolmogorov-smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
52. Lin, J. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
53. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
54. Mallows, C. L. A note on asymptotic joint normality. *Ann. Math. Stat.* **43**, 508–515 (1972).
55. Holme, P. Modern temporal network theory: a colloquium. *Eur. Phys. J. B* **88**, 1–30 (2015).
56. Anderson, R. M. & May, R. M.*Infectious diseases of humans: dynamics and control* (Oxford Science Publications, 1991).
57. Moinet, A., Starnini, M. & Pastor-Satorras, R. Burstiness and aging in social temporal networks. *Phys. Rev. Lett.* **114**, 108701 (2015).
58. Moinet, A., Starnini, M. & Pastor-Satorras, R. Aging and percolation dynamics in a non-poissonian temporal network model. *Phys. Rev. E* **94**, 022316 (2016).
59. Gelardi, V., Le Bail, D., Barrat, A. & Claidiere, N. From temporal network data to the dynamics of social relationships. *Proc. R. Soc. B* **288**, 20211164 (2021).
60. Stehlé, J., Barrat, A. & Bianconi, G. Dynamical and bursty interactions in social networks. *Phys. Rev. E* **81**, 035101 (2010).
61. Rocha, L. E. & Blondel, V. D. Bursts of vertex activation and epidemics in evolving networks. *PLoS Comput. Biol.* **9**, e1002974 (2013).
62. Vestergaard, C. L., Génois, M. & Barrat, A. How memory generates heterogeneous dynamics in temporal networks. *Phys. Rev. E* **90**, 042805 (2014).
63. Zhao, K., Stehlé, J., Bianconi, G. & Barrat, A. Social network dynamics of face-to-face interactions. *Phys. Rev. E* **83**, 056109 (2011).
64. Zhang, X., Moore, C. & Newman, M. E. Random graph models for dynamic networks. *Eur. Phys. J. B* **90**, 1–14 (2017).
65. Vanhems, P. et al. Risk of influenza-like illness in an acute health care setting during community influenza epidemics in 2004-2005, 2005-2006, and 2006-2007: a prospective study. *Arch. Intern. Med.* **171**, 151–157 (2011).
66. Swiderski, B., Osowski, S., Kruk, M. & Kurek, J. Texture characterization based on the kolmogorov–smirnov distance. *Expert Syst. Appl.* **42**, 503–509 (2015).
67. Baselice, F., Ferraioli, G., Pascazio, V. & Sorriso, A. Denoising of mr images using kolmogorov-smirnov distance in a non local framework. *Magn. Reson. imaging* **57**, 176–193 (2019).
68. Zierk, J. et al. Reference interval estimation from mixed distributions using truncation points and the kolmogorov-smirnov distance (kosmic). *Sci. Rep.* **10**, 1704 (2020).
69. Lopes, R. H., Reid, I. & Hobson, P. R. The two-dimensional kolmogorov-smirnov test. XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research. (2007).
70. Luiz, A. J. B. et al. Application of the kolmogorov-smirnov test to compare greenhouse gas emissions over time. *Braz. J. Biometrics* **39**, 60–70 (2021).

Acknowledgements

This research was supported by TAILOR, a project funded by the EU Horizon 2020 research and innovation program under GA No 952215. A.L., A.P., and B.L. acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU. B.L. and A.P. also acknowledge the support of the project AI@Trento (FBK-Unitn). G.C. acknowledges the support of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101103026 and from the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU.

Author contributions

The conceptualization of the idea was initiated by A.L. and G.C. Subsequently, A.L. undertook the development of the software implementation, while A.L. and G.C. jointly conducted the experiments under the supervision of A.P. and B.L. The manuscript was collaboratively written by G.C., S.L., A.P. and B.L, with project supervision provided by A.P. and B.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-023-01517-1.

Correspondence and requests for materials should be addressed to B. Lepri.

Peer review information *Communications Physics* thanks Sergey Shvydun and Alexandre Bovet for their contribution to the peer review of this work. A peer review file is available.