International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-022-00375-4

REGULAR PAPER

Δ-Conformity: multi-scale node assortativity in feature-rich stream graphs

Salvatore Citraro[1,2] · Letizia Milli[2] · Rémy Cazabet[3] · Giulio Rossetti[2]

Received: 30 November 2021 / Accepted: 8 November 2022

Abstract
Multi-scale strategies to estimate mixing patterns are meant to capture heterogeneous behaviors among node homophily, but they ignore an important *addendum* often available in real-world networks: the time when edges are present and the time-varying paths that edges form accordingly. In this work, we go beyond the assumption of a static network topology to propose a multi-scale, path- and time-aware node homophily estimator specifically tied for feature-rich stream graphs: Δ-*Conformity*. Our measure can capture the homogeneous/heterogeneous tendency of nodes' connectivity along a period of time Δ starting from a given moment in time. Results on face-to-face interaction networks suggest it is possible to track changes in social mixing behaviors that coincide with contextually reasonable everyday patterns, e.g., medical staff disassortative behavior when exposed to patients. In a different domain, that of the Bitcoin Transaction Network, we capture relationships between the quantity of money sent from (and to) different categories/continents and their respective mixing trends over time. All these insights help us to introduce Δ-*Conformity* as a suitable solution for understanding temporal homophily by capturing the mixing tendency of entities embedded in fine-grained evolving contexts.

Keywords Mixing patterns · Homophily · Dynamic networks · Stream graphs

1 Introduction

Networks help scientists to represent phenomena of uncountable complexity. The *network model*—in its minimal definition of a set of nodes with edges linking them—is acknowledged by researchers from many domains, leading to the consolidation of a multidisciplinary field crossing over math and physics [23], economy and social sciences [7], and any future productive contamination [6]. While analyzing complex data, the network topology is one of the most informative characteristics to focus on, making network science a solid paradigm for unveiling universal principles from different and domain-specific systems [27]. Nevertheless, a paradigm evolves as its research questions evolve: Nowadays, increasingly new approaches aim to combine the expressive power of topology to all those domain-specific aspects often available from complex systems. From attributes describing nodes to the insightful addendum of the temporal dimension, such enriched aspects can enhance valuable knowledge hidden in complex systems. The augmented structures incorporating them are often referred to as *feature-rich networks* [11]. In this work, we focus on specific instances of such augmented topologies: *node-attributed* and *dynamic* graphs. The former aspect focuses on the study of networks whose nodes are semantically enriched by context dependent attributes, the latter on frameworks designed to analyze the evolution of complex systems over time.

Leveraging such enriched network models, we study the evolution of *homophilic/heterophilic* behaviors—a critical emerging behavior in social systems—over time. Indeed, network science literature has deeply investigated homophily by introducing several measures to quantify it in networks, but little attention was given to its relation with the tem-

✉ Salvatore Citraro
salvatore.citraro@phd.unipi.it

Letizia Milli
letizia.milli@isti.cnr.it

Rémy Cazabet
remy.cazabet@univ-lyon1.fr

Giulio Rossetti
giulio.rossetti@isti.cnr.it

1 Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo, 3, 56125 Pisa, Italy

2 KDD-Lab, CNR - ISTI, Via G. Moruzzi, 1, 56125 Pisa, Italy

3 LIRIS, Université Lyon 1 - CNRS, Lyon 69622, France

Published online: 27 November 2022

poral dimension—i.e., studying to what extent it remains stable/changes as the underlying topology changes. Being interested in tracking nodes' homophily over time, we leverage the combined expressive power of node-attributed and dynamic graphs—these latter ones defined through the formalism of stream graphs [16]. In this work, we propose a node-centric, path-aware, homophily measure for attributed stream graphs, i.e., Δ-*Conformity*, as a conservative extension of *Conformity* [31], a recent multi-scale measure aiming to capture the heterogeneity of mixing patterns in networks. Δ-*Conformity* leverages the concept of time-respecting paths, and it is able to cope with both static and varying categorical attributes.

The rest of the paper is organized as follows. Section 2 introduces the main literature needed to understand the current work fully; Sect. 3 introduces and formally describes Δ-*Conformity*; Sect. 4 provides several case studies on which the proposed formula can be applied; Sect. 5 concludes the work.

2 Related works

An overview of several topics is provided here, namely (i) homophily estimation in node-attributed static graphs, (ii) dynamic networks representations, and (iii) homophily definitions in dynamic environments.

2.1 Homophily estimation

In social networks, *homophily* refers to the tendency of people to be more likely to interact with similar others w.r.t. several social dimensions, from age to political leaning, often grounded by households, workplaces, geographical environments. [19]. Studying homophilic behavior may help researchers to unveil critical networks dynamics, as in the studies of (i) segregation in interracial friendships [21], (ii) gender-specific patterns in early school grades [34], (iii) sexually transmitted diseases spreading [2], (iv) trustworthiness in business networks [3].

The Newman's *assortativity coefficient* r [23] is a popular measure that quantifies homophily in complex networks. This quantity is calculated as the sum of the differences between the observed and the expected fraction of edges between nodes sharing similar values of an attribute. When maximized, $r = 1$, the assortativity coefficient describes a network where all edges connect to nodes labeled with the same value; when $r = 0$, the edges are randomly connected; when minimized, $r = -1$, the coefficient describes a network where all edges connect to nodes with a different value. Other assortativity measures, like *ProNe* [28], or statistical approaches, like the *VA-Index* [26], include more complex scenarios, estimating the correlation between structure and two or more attributes.

Nevertheless, a limitation of these measures is that they sum up a single, average network behavior, ignoring/absorbing the plausible existence of outliers and heterogeneous mixing patterns. Considering individual connectivity preferences can shed light on important network dynamics, as in the studies on (i) *monophily* [1] (only some individuals show preferences for labels unrelated to their own); (ii) *perception biases* [17] (homophily combined to the minority size of a group are responsible for false consensus or false uniqueness). Inferring and quantifying individual differences comes as a hard task in complex network analysis. In presence of strong variations, the mean value of a group is unable to fully describe individual node preferences [5]. Hence, several works focused on multi-scale strategies to estimate homophily. In [25], node similarity is measured as the autocorrelation of a time-series defined as a sequence of node labels visited by a random walker allowing a restart. The assumption is that random walks can integrate information about paths of all possible lengths, extracting similarities from a higher context than the adjacent neighborhood only. A similar approach is used in [9], applied in graph classification: Multi-hop assortativity is defined as the probability that a randomly selected node and a randomly selected t-hop neighbor belong to the same category, where t indicates the time of the visit of the random walker. In [4], the focus is on the mean first passage times between preassigned classes of nodes, i.e., the expected number of steps needed for a random walker to visit for the first time a node of a certain class when it starts from a node of another class. This concept is used to estimate nodes' heterogeneity, polarization and segregation.

2.2 Dynamics of networks

In complex networks, when time is involved, differentiating between persistent and instantaneous connections is fundamental. Friendships relations and scientific collaborations are persistent over time, whereas phone calls and physical proximity interactions can only have a certain duration. Emails, messages on social media or financial transactions are inherently instantaneous. Hence, several models are needed to properly represent all these different dynamics. The most used representations are the Graph Snapshot (SN), the interval Graph (IG), and the Link Stream (LS). In SNs, a network is represented as a sequence of graphs, each one analyzed autonomously and independently from the others. In IGs, the focus is on the characterization of link durations, these ones identified by a start and an end in time. In LSs, a link is identified only by a pair of nodes and an instantaneous point in time. Each network representation has its *pros* and *cons*. SNs can capture significant information but can lead to losses of temporal information, e.g., paths within a

snapshot do not respect the time-varying dynamics of interactions. IGs allow to define time-respecting paths, where the focus is on the nodes that can be reached from other ones within some observations window; however, choosing the appropriate time window may not be trivial. More than the others, LSs consider the streams of interactions over time, taking into account both the temporal and the structural nature of interactions. Recent works tried to further generalize such representations: For instance, the Stream Graph model [16] aims to create a formalism dealing with both instantaneous links and links with duration. Stream graphs can easily extend and generalize static centrality measures as the betweenness [37], and analyze empirically the differences in shortest, fastest, foremost time-respecting paths [35,36]. Recently, augmented stream graphs models have been defining, as in modeling interactions over time with multilayer structure [24], where the focus is on the definition of new layer centrality measures.

2.3 Toward temporal homophily estimation

The temporal dimension of homophily is still little understood in complex network analysis. Introducing new dimensions in homophily estimation is not trivial. For instance, a definition of anti-assortative and anti-disassortative mixing patterns is needed while studying homophily in trust or signed networks, where sharing similar values must be related to the positive or negative value of an interaction [29]. Introducing a temporal dimension means dealing with different time representations. Hence, temporal homophily can present different facets. Several works addressed different research questions, giving more facets and complexity to the problem. For instance, some works about degree-degree assortativity describe the evolution patterns of this property. In [40], a universal behavior in temporal homophily is explained as follows: *degree assortativity increases at the beginning of network evolution and decreases to a long-lasting stable level*. This behavior was observed independently in several domain-specific domains, as in the Bitcoin Transaction Network [14]. *Temporal homophily* has been defined in [15] by leveraging the notion of colored networks. Here, homophily is described as the tendency of similar nodes to participate in mixing patterns—through occurrences of node colors in temporal motifs—beyond what would be expected from a null model representing the structure of the aggregate network, assuring indeed that dynamic observations are independent of results obtained from a static analysis. The interplay between structure, attributes, and time has also been addressed in [33] as a relation between a first-order and second-order similarity, where the former one is the common definition of homophily (between attributes sharing similar values), and the latter one is the similarity in trajectories of changes.

3 Δ-Conformity

Our work aims to propose a measure able to characterize homophilic behaviors in presence of time evolving topologies. During the last decade, several approaches have been proposed to support dynamic network analysis; in this section, we firstly provide a description of the modeling paradigm adopted, then describe the *Conformity* measure and its extension—namely Δ-*conformity*—to time evolving networks.

3.1 Feature-rich stream graphs

Two peculiar features characterize the network structures whose homophily we are interested in measuring: (i) their nodes are enriched by categorical attributes and, (ii) their node and edge's sets are allowed to vary as time goes by. Since such information can be modeled leveraging several different frameworks, it is mandatory to properly define the reference framework used. For the sake of simplicity, we introduce our model—Feature-rich Stream Graph—through incremental definitions, introducing first node annotations and then temporal dimension.

When in the presence of a static network whose nodes exposes semantic properties, we define a Node-attributed graph as

Definition 1 (*Node-attributed Graph*) $\mathcal{G} = (V, E, L)$ is a node-attributed graph, where V is the set of nodes, E is the set of edges, and L is a set of categorical attributes such that $L(v)$, with $v \in V$, identifies the set of categorical values associated to v.

Stream graph [16] is a powerful modeling formalism that allows us to describe and characterize streams of nodes and edges having an associated lifespan (i.e., appearing and vanishing, even multiple times, during the network life). Therefore, we leverage this mature formalism to integrate the temporal dimension within the Node-attributed graph framework:

Definition 2 (*Feature-rich Stream Graph*) $\mathcal{S} = (T, V, W, E, L)$ is a stream graph, where $T = [\mathrm{A}, \Omega]$ is the set of discrete time instants, with A and Ω the initial and final instants, $W \subseteq T \times V$ the set of temporal nodes, $E \subseteq T \times V \times V$ the set of edges such that $(t, uv) \in E$ implies $(t, u) \in W$ and $(t, v) \in W$ and L the set of temporal node attributes such that $L(t, v)$ with $v \in V$ and $t \in T$ identifies the set of categorical values associated to v at time t.

Feature-rich stream graphs allow describing those networks whose topology and node attributes vary as time goes by. Consider the example in Fig. 1, where edges and their duration are identified, respectively, by vertical and solid

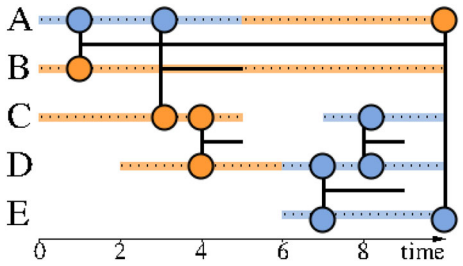


Fig. 1 A feature-rich stream graph with time-varying labels. Vertical and solid horizontal segments identify edges and their duration, respectively; dotted horizontal segments are the node labels, e.g., A's label from $t = 0$ to $t = 5$ is blue, changing in orange from that point on (color figure online)

horizontal segments, while node attributes are represented by node colors; a node maintains a label along the dotted horizontal lines until the color changes. To contextualize the toy example, we can imagine the edge identifying pairwise discussions among users in an online social media, while node attributes identify individual stances on a given topic (e.g., the left/right leaning of the users involved in a political debate).

3.2 Multi-scale homophily in feature-rich stream graph

We describe Δ-*Conformity* as a conservative extension of *Conformity* (previously introduced in [31]) to Feature-rich Stream Graph.

Given a node-attributed (static) graph, *Conformity* is a multi-scale homophily measure that accounts for the length of paths connecting node pairs while computing individual degrees of assortativity. To better clarify how *Conformity*—and by extension Δ-*Conformity*—works, we provide its formalization.

Definition 3 (*Conformity*) Given a real number α in $[0, +\infty)$, the *Conformity* score for a node $u \in V$ is defined as

$$\psi(u, \alpha) = \frac{\sum_{d \in D} \frac{\sum_{v \in N_{u,d}} I_{u,v} f_{v,l_v}}{|N_{u,d}| d^{\alpha}}}{\sum_{d \in D} d^{-\alpha}}, \tag{1}$$

where

- D is the maximum distance among all node pairs $(u, v) \in V$;
- $N_{u,d}$ is the set of u's neighboring nodes at distance d, i.e., the nodes that can be reached after d number of hops from the target node u;
- α tunes the relevance of nodes at distance d by the source u;
- $I_{u,v}$ is the indicator function comparing the attribute values of a node u and a node v

$$I_{u,v} = \begin{cases} 1 & \text{if } l_u = l_v \\ -1 & \text{otherwise,} \end{cases} \tag{2}$$

- f_{u,l_u} is the similarity function computing the ratio of u's first-order neighbors (namely $\Gamma(u)$) sharing its same attribute value l_u—a ratio forced to the range in (0, 1] by imposing its value to 1 when the numerator nullifies

$$f_{u,l_u} = \frac{|\{v | v \in \Gamma(u) \wedge l_u = l_v\}|}{|\Gamma(u)|}. \tag{3}$$

In the measure, we compute the similarity function on each u's neighbor v, namely f_{v,l_v}.

The final score is thus normalized to ensure that *Conformity* lies in the range $[-1, 1]$: the lower bound identifying heterophilic behaviors, the upper homophilic ones.

In its original definition [31], the distance function adopted by *Conformity* is the one computing network geodesic paths. To cope with temporal constraints introduced by dynamically evolving topologies, such a choice needs to be revised.

While adopting the Feature-rich Stream Graph model, we are particularly interested in observing interactions occurring at least once every Δ units of time—where Δ is a time interval of a certain duration in T. Therefore, we define Δ-*Conformity* as a Δ-analysis [16] approach:

Definition 4 (Δ-*analysis: Stream Graph*) $\mathcal{S}_\Delta = (T_\Delta, V, W_\Delta, E_\Delta, L_\Delta)$ is the attributed stream graph such that $T_\Delta = [\mathrm{A} + \Delta, \Omega - \Delta]$, an edge $(t', uv) \in E_\Delta$, with $u \in W_\Delta$ and $v \in W_\Delta$, is present at a time t' in $\mathcal{S}_\Delta$ and each node $v \in V$ has a set of categorical values $L(t', v)$ whenever it is present in $\mathcal{S}$ at a time t in $[t' + \Delta]$.[1]

Once fixed such a reference framework, we can define Δ-*Conformity*:

Definition 5 (Δ-*Conformity*) Given a real number α in $[0, +\infty)$, a temporal id $t \in T$, and a time window Δ, the Δ-*Conformity* score for a node $u \in V$ is defined as the value of Conformity(α, u) where

- the Feature-rich Stream Graph is restricted to $\mathcal{S}_\Delta = (T_\Delta, V, W_\Delta, E_\Delta, L_\Delta)$ with $T_\Delta = [t, t + \Delta]$;
- the distance among $u, v \in V$ is the length of a time-respecting path connecting the two (if it exists, ∞ otherwise).

[1] The interval considered in [16] is $\left[t' - \frac{\Delta}{2}, t' + \frac{\Delta}{2}\right]$; in our formalism, we remain consistent with the analyses performed in the experimental section.

The proposed definition introduces three peculiarities that make Δ-*Conformity* more complete w.r.t. its predecessor: (i) the concept of *memory* (modeled with Δ), (ii) of *preferential* temporal interaction patterns (made explicit by the distance function) and, (iii) the new role of α, as tuning parameter controlling the relative importance of elements within a time-respecting path w.r.t. their distance in time from the target node (the higher the value of α the smaller the contribution to the score of nodes that are reachable farther in time from the source). In particular, while Δ specifies the maximum time span for a path (and, thus, its maximum length), α allows tuning each node relative importance given their position in the identified paths.

The former aspect defines the temporal bounds for computing homophily: The underlying idea is that the communication chains used to evaluate nodes' similarity need to be bound to a specific temporal window (namely, as time pass new interactions in the chain lose their relevance for the source node). The latter aspect instead focuses on how such chains are built. If in static networks the simplest function to compute distances among two nodes is computing the length of the shortest path connecting them, several alternative definitions can be used to reach a similar goal in dynamic networks.

Indeed, the notion of *distance* is crucial while emphasizing the dynamic nature of graphs. Paths in stream graphs have both a length and a duration. A time-respecting path can be defined as in the following:

Definition 6 (*Time-respecting path*) In a stream graph $S = (T, V, W, E)$, a sequence $(t_0, u_0, v_0), \ldots, (t_k, u_k, v_k)$ of elements $T \times V \times V$, such that $u_0 = u$, $v_k = v$, $t_o \geq \alpha$, $t_k \leq \omega$, for all i, $t_i \leq t_{i+1}$, $v_i = u_{i+1}$, and $(t_i, u_i, v_i) \in E$, $[\alpha, t_0] \times u \subseteq W$, $[t_k, \omega] \times v \subseteq W$, and for all i, $[t_i, t_{i+1}] \times v_i \subseteq W$ is a time-respecting path from $(t_0, u_0) \in W$ to $(t_k, v_k) \in W$ of length k and duration $t_k - t_0$.

Time-respecting paths are computed on the temporal DAG (directed acyclic graph) rooted in a source node u at time t—where each edge identifies an interaction among two nodes if they are reachable in contiguous snapshots. By definition, such induced topology does not allow the existence of paths composed of directed edges occurring at the same time (e.g., it is not possible to have a path $[(i, j, 0), (j, z, 0]$ from i to z). From such observation, it is derived that, although we can compute for each node a time-series of Δ-Conformity scores, this does not imply that each of such scores is the result of a snapshot analysis; rather that it represents the homogeneity/heterogeneity tendency of that node's connectivity for a period of Δ starting from a given moment in time.

Indeed, the time-respecting path definition naturally leads to different notions for the cost of reaching nodes, depending on specific paths constraints. As an example, considering a Stream Graph $S = (T, V, W, E)$, we can focus on

- *Shortest path* $\mathcal{P}$ is a shortest path from $(t_0, u_0) \in W$ to $(t_k, v_k) \in W$ if it has minimal length k;
- *Fastest path* $\mathcal{P}$ is a fastest path from $(t_0, u_0) \in W$ to $(t_k, v_k) \in W$ if it has minimal duration $t_k - t_0$;
- *Foremost path* $\mathcal{P}$ is the path from $(t_0, u_0) \in W$ to $(t_k, v_k) \in W$ that, independently from its length and duration, allows to reach first the destination.

While used within the framework of Δ-*Conformity*, those three path types—that represent only a subset of interesting ones—can profoundly affect the observed homophily/heterophily of individual nodes.

Consider for instance the toy example reported in Fig. 2 involving the temporal interactions among five nodes $V = [A, B, C, D, E]$ in $T = [0, 10]$. There we assume $\Delta = 5$ and compare shortest and foremost paths starting from A, namely:

- Shortest paths:

  (A,B) = [(A,B,0)]
  (A,C) = [(A,C,2)]
  (A,E) = [(A,E,4)]

- Foremost paths:

  (A,B) = [(A,B,0)]
  (A,C) = [(A,B,0),(B,C,1)]
  (A,E) = [(A,C,2),(C,E,3)]

Considering two possible node labels, blue and orange, assigned, respectively, to the node sets (A, C, E) and (B, D), we can observe that Δ-*Conformity* ($\alpha = 1$) unveils completely different homophilic patterns for node A: assortative while considering shortest paths, disassortative for foremost paths.

Such a counterintuitive behavior is tied to the different distances that the two time-respecting paths generate, and its interpretation can be easily clarified, providing more semantic to our example. Let us assume that the previous stream graph models a word of mouth-like diffusion phenomenon and that users' labels describe opposing stances regarding the discussed topic: Every time the content is passed to a peer having a different opinion from the source, its original value is reduced. Foremost paths aim to reach all the available users as early as possible, disregarding that the original content might reach same-stance users only after being dismantled piece after piece. On the other hand, shortest paths aim to minimize such an effect by cutting down the length of message passing chains.

Indeed, one of the advantages of Δ-analysis lies in the opportunity of analyzing the temporal trends of a given indicator. For Δ-*Conformity*, this means observing—node-wise—how homophilic/heterophilic behaviors unfold as

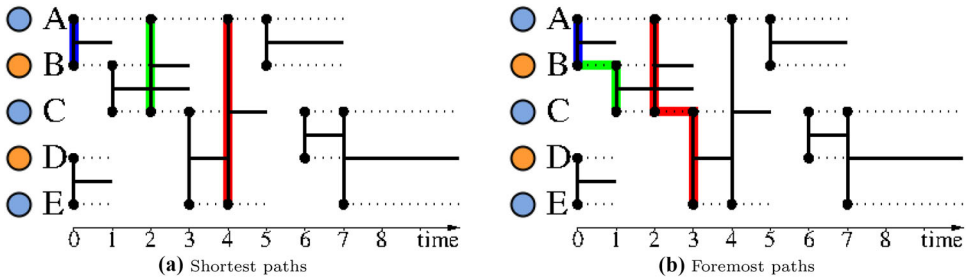


Fig. 2 Time-respecting paths. A comparison of shortest (**a**) and foremost (**b**) paths starting from the node A, at $t = 0$, and targeting nodes reachable within a time window $\Delta = 5$. Assuming the set blue $= (A, C, E)$ and orange $= (B, D)$ of nodes sharing a same label and applying the proposed measure, we observe that: **a** shortest paths identify an assortative behavior $\Delta\text{-}Conformity(A) = 0.33$ while **b** foremost ones identify a disassortative one, $\Delta\text{-}Conformity(A) = -0.41$ (color figure online)

time goes by. Leveraging such rationale, in the following section, we underline how peculiar Δ-*Conformity* trends can characterize different actors of real-world evolving complex systems.

4 Experiments

In this section, we explore Δ-*Conformity* testing the measure on several real-world datasets from different domains. Table 1 sums up the main characteristics of the chosen datasets, such as the aggregated number of nodes and edges as well as common dynamic network statistics used for stream graphs analysis [16].

We focus on finding heterogeneous mixing behavior by studying the characteristics of single nodes or classes of nodes over time. We also focus on testing the statistical significance of the emerging trends, i.e., whether they are interesting patterns of the datasets or whether distributions from randomized networks let similar patterns emerge. In particular, this quantitative analysis is performed on the Bitcoin Network (cf. *Bitcoin Network*).

In all the experiments, we focus on the *shortest paths* only, fixing the decay parameter α to 2 [31]. Different values of Δ are used with respect to the analysis purposes or the dataset nature.

4.1 Copenhagen network study

We consider the 700 (male and female) university students participating in the Copenhagen Network Study [32], whose social interactions are estimated via Bluetooth signal strength. In [31], *Conformity* by gender was measured in a daily aggregated static network, where males were on average more assortative than females. We aim to provide more insights into students' daily routines, e.g., whether mixing patterns differ by day and night or by weekdays and weekends. Hence, we consider a dynamic network covering little more than 1 week, i.e., from Sunday 12 am to Tuesday 2 pm. Timescale aggregation window is by *hours*, then we consider two different Δ, 1 and 8, where interactions and paths occur within a window of 1 or 8 hour(s). Figure 3 reports the average Δ-*Conformity* trends of male and female students. Similar to the static aggregated analysis [31], females present a disassortative trend, while male trends are assortative. Interestingly, the average values differ when using two different values of Δ. Using larger windows implies including more links in the analysis, i.e., higher average degree $\langle k \rangle$. This can explain why the average Δ-*Conformity* scores are closer to the horizontal lines: The higher the aggregation of links, the lesser the distinctiveness of groups' mixing patterns, where Δ-*Conformity*=0 corresponds to the uniformly mixed behavior. However, larger windows let emerge periodicity. In particular, the trend of male students correlates with the average degree trend of the group with Δ=8. Thus, higher values of Δ can suggest a broad point of view that can capture the circadian nature of people interactions, and both the two network indicators, i.e., Δ-*Conformity* and average degree $\langle k \rangle$, can help to describe these behaviors better.

4.2 SocioPatterns

We consider three datasets from the SocioPatterns collection: *Hospital ward* [39] is a set of contacts between patients and health-care workers; *Primary school* [38] and *High school* [18] are two sets of contacts and friendship relations between children/high school students.[2] We focused on one day in all three networks. Timescale aggregation window is by *hours*. We set $\Delta = 1$ to capture fine-grained activities, e.g., lunch breaks. Figure 4 sums up the analysis on these networks.

The first panel—Fig. 4a—introduces the hospital ward dataset. Black lines represent the average Δ-*Conformity* trends of the groups, while colored points are individual nodes, and the corresponding dotted colored lines follow

[2] http://www.sociopatterns.org/.

Table 1 Dynamic network statistics for all the datasets, where $|V_G|$ and $|E_G|$ are the number of nodes and edges, respectively, in the aggregated graph, $|T|$ is the number of time instants, $|V|$ is the number of nodes (henceforth, in the stream graph), $|W|$ is the number of temporal nodes, $|E|$ is the number of edges, $cov.$ is the stream graph coverage [16], and L is the set of attribute names (attribute cardinality in brackets)

	$\|V_G\|$	$\|E_G\|$	$\|T\|$	$\|V\|$	$\|W\|$	$\|E\|$	$cov.$	L
Copenhagen	633	33622	216	191.89	65.48	155.65	0.30	Gender (2)
Hospital ward	49	460	16	25.06	8.18	28.75	0.51	Work category (4)
Primary school	47	324	10	21.80	4.63	32.40	0.46	Gender (2), Class (4)
High school	180	758	9	93.33	4.66	84.22	0.51	Gender (2), Class (4)
Bitcoin network	84	2275	335	38.18	152.27	6.79	0.45	Category (4), Continent (5)

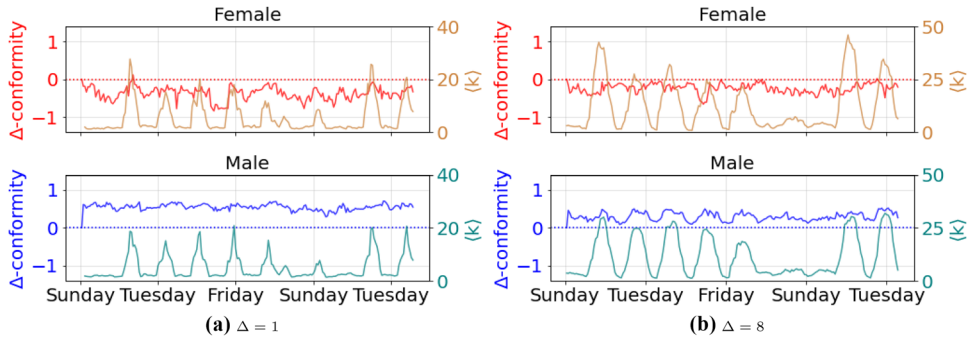


Fig. 3 Copenhagen network study. Δ-*Conformity* by gender using **a** $\Delta = 1$ and **b** $\Delta = 8$; the two *y*-axes indicate the average Δ-*Conformity* score of the group (left axis) and the average degree $\langle k \rangle$ of the group (right axis) of the category. The horizontal colored lines, i.e., Δ-*Conformity* = 0, highlight the uniformly mixed behavior (color figure online)

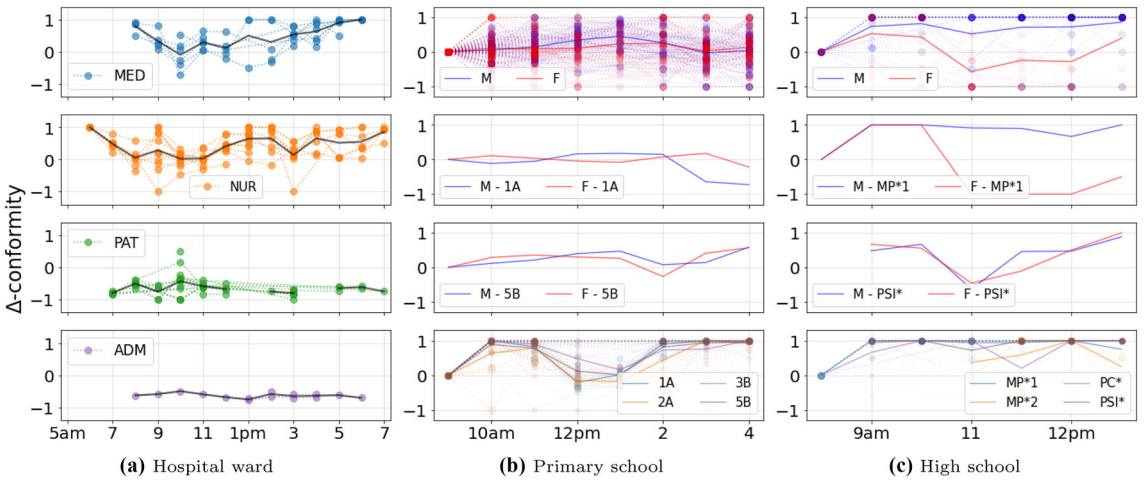


Fig. 4 SocioPatterns. **a** Δ-*Conformity* by hospital category; **b, c** from top down: Δ-*Conformity* by gender: trends showing the whole network (first subplot) and the subgraphs of two distinct classes (second, third subplots); Δ-*Conformity* by class (last subplot); **a** black or **b, c** colored lines represent the average value of the groups, while points are the values of individual nodes (color figure online)

each node for tracking the evolution of the score (medical doctors in blue, nurses and nurse' aids in orange, patients in green and administrative staff in purple). The analysis starts from $6am + \Delta$, when contacts are between nurses and nurse' aids only; this results in perfectly assortative group behavior. When nurses and nurse' aids start to visit patients [39] ($7am + \Delta$), nodes' mixing splits into two branches, one exhibiting assortativity and the other one disassortativity. Hence, individual differences become visible, and the average Δ-*Conformity* score of the group, i.e., the black line, is no longer useful for capturing group heterogeneity. On average, patients remain disassortative all day; since *all rooms but 2 were single-bed rooms* [39], also the assortative behavior of some patients at $10am + \Delta$ is explained. The absence of patients at certain times of the day can be motivated with the lunch break (e.g., $1pm + \Delta$) or with a pause of the visit time (e.g., $4pm + \Delta$). The few administrative staff members are disassortative all the time.

The other two panels—Fig. 4b, c—introduce the analysis on the two schools. We measure homophily by gender and homophily by class. From top down, the first three subplots (of both schools) focus on homophily by gender. The first

subplot focuses on the whole network visualization, while the other two focus on two selected classes' subgraphs. Colored solid lines represent the average trends of the groups, while points are the scores of individual nodes. Interestingly, primary and high school students behave differently with respect to the gender attribute. Primary students are uniformly mixed. Conversely, the behavior of high school students is similar to what we have already observed in the Copenhagen college network: Males are on average more assortative than females. Focusing on a subset of classes (second, third subplots) allows us to explain in which classes this divergent behavior among groups is stronger; e.g., mathematics and physics students (*MP*1*) [18] present more differences than engineering students (*PSI**). The last subplot of each school shows Δ-*Conformity* by class. As expected, high schools students are strongly assortative with respect to the class, motivated by the fact that students of different classes do not know each other. A more interesting pattern occurs in primary school. In line with the data described in the reference paper [38], the disassortative trends occurring from $12am + \Delta$ to $1pm + \Delta$ can be explained through the children spatio-temporal trajectories: Children move from their rooms to the playground or cafeteria for lunch; hence, it is more likely for a child to enter in contact with children of another class.

4.3 Bitcoin network

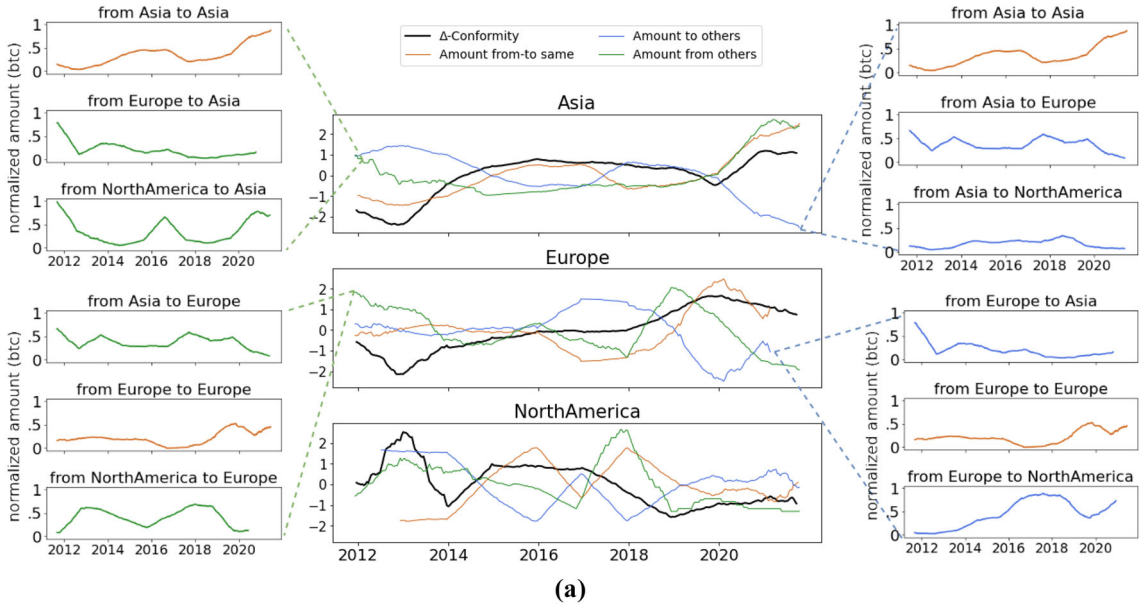
We consider the network of Bitcoin transactions extracted from the blockchain between blocks 0 and 667542 (January 2021). Actors (groups of addresses) are identified using the standard co-input heuristic [10,30], and we create a daily aggregated network in which there is an edge between two actors if there is an observed transaction between them on that day. We filter the network to keep only the top 100 actors with the most transactions (cf. later). Similarly to [13], we label each user with its *category* from the WalletExplorer,[3] among Exchanges, Gambling, Pools, and Services. Moreover, we label each node with an attribute identifying actor's headquarter *location*, among Asia, Australia, Central & South America, Europe, North America, and those illegal websites without a physical location labeled as *Dark Web*. After the labeling operation, we filter out further nodes whose we did not find the proper category, hence we finally keep 84 nodes out of 100. Each link in the network is weighed with the amount of Bitcoin sent to/from each node. Assuming an undirected network while estimating node Δ-*Conformity*, we distinguish between the amount sent to/from each group only for adding more insights in the interpretation of Δ-*Conformity* trends. Remember that *Conformity* does not consider link weights in its measurement.

We study the two attributes separately. We consider Δ=1, hence links occurring day by day. Figures 5and 6 sum up the analysis on Δ-*Conformity* by location and by category, respectively. In both figures, Δ-*Conformity* trends are plotted together with the amount of Bitcoin sent to and from each location and category, respectively—Figs. 5a, 6a. Groups' amounts are normalized w.r.t. the total amount of money sent in each Δ window considered. This information is shown in detail in the subplots on the left and right sides of the panel: left, amount sent from the other groups to the target one; right, amount sent from the target group to the other ones. For better readability, in Fig. 5a, we only show the disagreggated trends between Asia, Europe and North America actors, which are also the most statistically significant ones according to their z-score trends (cf. later, Fig. 5c); the same reason—readability—applies to Fig. 6a, where only trends between Exchanges, Gambling and Services are shown: It is also evident here how the quantity of money sent from and to Exchanges and Services depends very much on each other and little on other categories. More in detail, Asian actors tend to send less money to other locations and bigger quantities between each other; similarly, European actors tend to send more money among themselves, but other locations send less money to them. Δ-*Conformity* trends seem to follow the same pattern, e.g., they are more assortative when the amount sent from-to the same attribute value are higher. Similarly, Exchanges actors mainly play among themselves, and their average Δ-*Conformity* trend is assortative on average, while Services actors mainly play with Exchanges actors, and their average Δ-*Conformity* trend is disassortative. The relationship between the Δ-*Conformity* trend and the quantity of money sent between the same actor typology is a preliminary qualitative validation of our measure, in particular when a higher quantity of money sent from and to similar actors is related to a lower quantify of money sent from and to other categories (see Asia).

While aligning Δ-*Conformity* and Bitcoin trends in one plot, we lose the original scale of Δ-*Conformity* scores. Hence, Δ-*Conformity* scores are also plotted in Figs. 5b and 6b, lying in the original range of values. Locations are slightly disassortative or uniformly mixed over time—Fig. 5b. Categories are mainly disassortative, except for Exchange, that are on average slightly assortative—Fig. 6b.

A value of Δ-*Conformity* below zero indicates that actors of the corresponding category do not interact mostly with actors of the same category, i.e., they do not form a *closed club*. However, they might still be interacting more frequently with actors of the same category than expected at random. To check this hypothesis, we use a bootstrap approach and z-scores to check if the Δ-*Conformity* is nevertheless significantly larger than expected by chance, thus pointing toward a preference for actors of the same category. Hence, panel (c) of both figures shows the z-scores trends assessing the

[3] https://www.walletexplorer.com/.



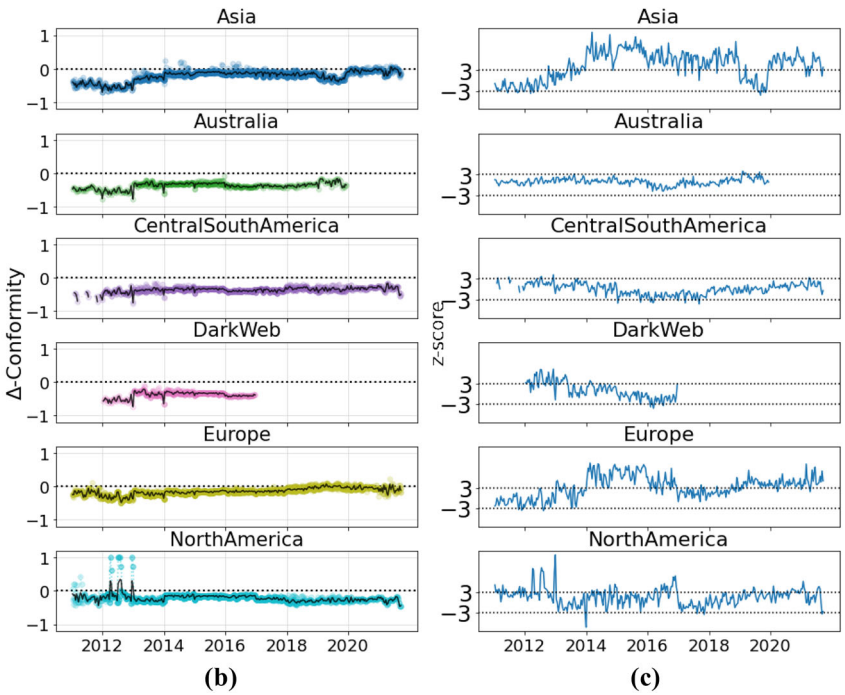


Fig. 5 Bitcoin transaction network—location attribute. **a** Comparison between the Δ-*Conformity* and the Bitcoin amount trends, with a focus on the amounts sent to/from Asia and Europe from/to other classes; **b** Δ-*Conformity* of different classes w.r.t. to the location attribute, and **c** their z-score trends

reliability of Δ-*Conformity* trends. Z-scores are obtained by comparing point by point the Δ-*Conformity* score of the original graph to a null distribution of 200 rewired networks using a configuration model [8,20]; we use the following formula for the comparison:

$$z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where x is the Δ-*Conformity* value in the original graph, μ is the average Δ-*Conformity* value from the ensemble of rewired networks, σ is the standard deviation of the ensemble, and n is the number of nodes having l as label, e.g., having Asia as location. Horizontal lines at values 3 and -3 are supporting to determine the statistical significance of original Δ-*Conformity* scores.

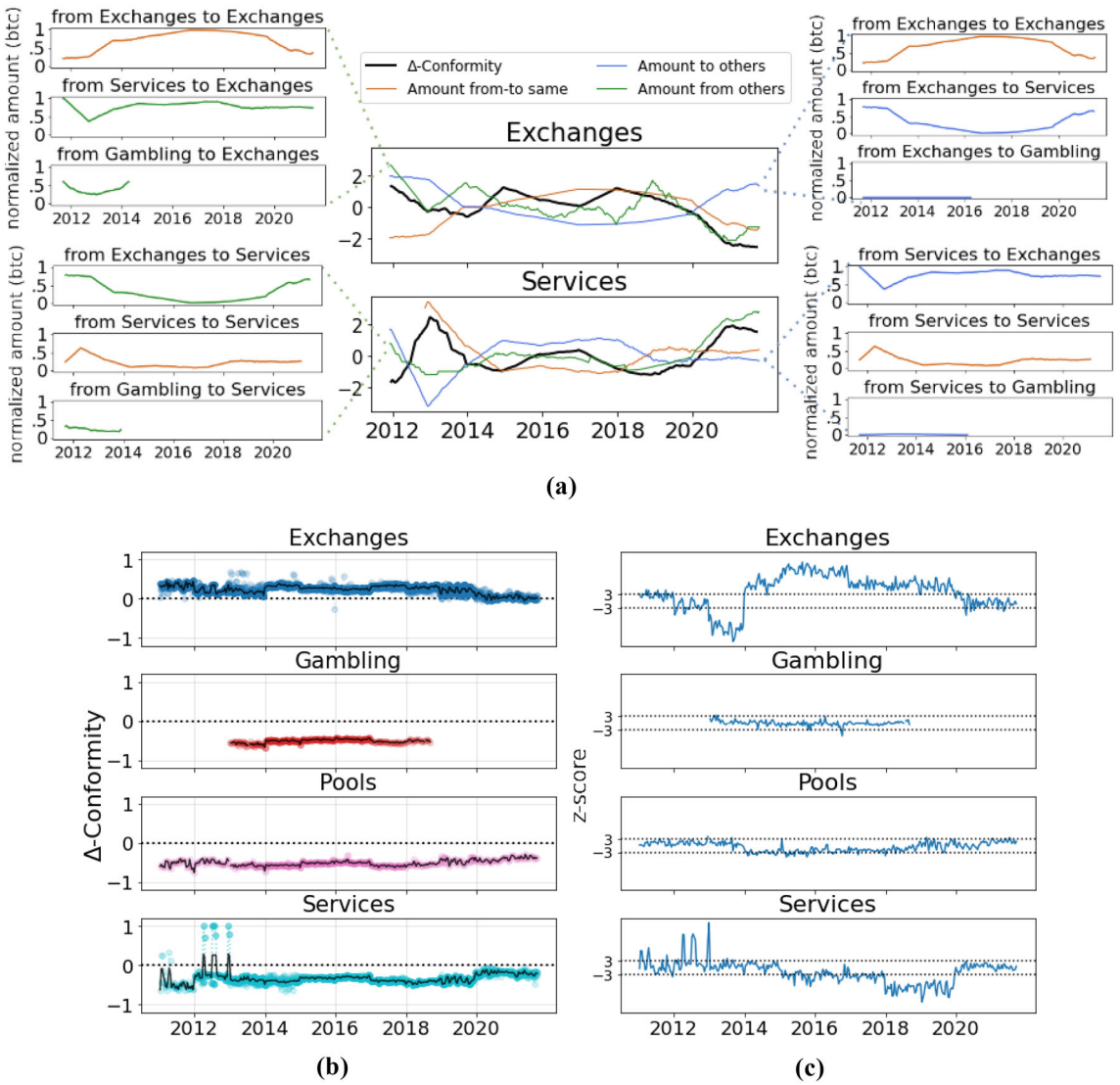


Fig. 6 Bitcoin transaction network—category attribute. **a** Comparison between the Δ-*Conformity* and the Bitcoin amount trends, with a focus on the amounts sent to/from Exchanges and Services from/to other classes; **b** Δ-*Conformity* of different classes w.r.t. to the category attribute, and **c** their z-score trends

We can suppose that the expected behavior of a rare/less frequent category is disassortative, because few of the encountered nodes are of the same category, whatever the distance. For instance, the very few *Australian* or *DarkWeb* nodes are highly disassortative, but this behavior is not significant. Conversely, the *Asian* nodes are still disassortative—they interact to each other and to other five classes as well—but their z-scores above 3 are meaning that this behavior is still higher than expected—cf. Fig. 5c. Similarly, Exchanges and Services Δ-*Conformity* trends are more statistically reliable than Gambling and Pools—Fig. 6c.

5 Conclusions

In this work, we explored node-centric mixing pattern estimation in feature-rich stream graphs by extending the recent multi-scale, path-aware measure of *Conformity* [31]. To lift the unrealistic assumption of a fixed network topology, we leveraged the formalism of the Δ-analysis of stream graphs [16]. Re-framing *Conformity* in such a framework allowed us to propose a more fine-grained and dynamic node-wise homophily estimator, i.e., Δ-*Conformity*, able to cope with time-varying paths, among shortest, fastest, foremost, etc., and with both static and varying node categorical attributes.

While analyzing Δ-*Conformity* trends in several social interaction networks, we found that the mixing behavior of node can change over time. Such changes coincide with con-

textually reasonable everyday patterns, from the working hours of medical staff to the lunch break of primary school children. Moreover, a broad case study on the Bitcoin Transaction Networks convinced us that Δ-*Conformity* could be applied on several and different domains, not only social networks.

As future works, we plan to apply Δ-*Conformity* for hot topics in computational social science, as in multi-scale echo-chamber identification [22], or for monitoring users in sensible online communities, e.g., mental health-related ones [12]. We expect that the same mechanisms applied on Bitcoin network could allow us to identify cliques of actors working together to manipulate the market or the formation of national markets, for instance in a country adopting Bitcoin as local tender such as El Salvador.

Acknowledgements This work is supported by the European Union—Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities", Grant Agreement no. 871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu) and by the CHIST-ERA grant CHIST-ERA-19-XAI-010, by MUR (grant No. not yet available), FWF (Grant No. I 5205), EPSRC (Grant No. EP/V055712/1), NCN (Grant No. 2020/02/Y/ST6/00064), ETAg (Grant No. SLTAT21096), BNSF (Grant No. КП-06-П-002/5). This work is partially supported by BITUNAM Project ANR-18-CE23-0004.

Funding Open access funding provided by Universitá di Pisa within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Altenburger, K.M., Ugander, J.: Monophily in social networks introduces similarity among friends-of-friends. Nat. Hum. Behav. **2**(4), 284–290 (2018)
2. Aral, S.O., Hughes, J.P., Stoner, B., Whittington, W., Handsfield, H.H., Anderson, R.M., Holmes, K.K.: Sexual mixing patterns in the spread of gonococcal and chlamydial infections. Am. J. Public Health **89**(6), 825–833 (1999)
3. Barone, M., Coscia, M.: Birds of a feather scam together: trustworthiness homophily in a business network. Soc. Netw. **54**, 228–237 (2018)
4. Bassolas, A., Nicosia, V.: First-passage times to quantify and compare structural correlations and heterogeneity in complex systems. Commun. Phys. **4**(1), 1–14 (2021)
5. Cantwell, G.T., Newman, M.E.J.: Mixing patterns and individual differences in networks. Phys. Rev. E **99**(4), 042306 (2019)
6. Coscia, M.: The atlas for the aspiring network scientist. arXiv preprint arXiv:2101.00863 (2021)
7. Gao, J., Zhang, Y.-C., Zhou, T.: Computational socioeconomics. Phys. Rep. **817**, 1–104 (2019)
8. Gauvin, L., Génois, M., Karsai, M., Kivelä, M., Takaguchi, T., Valdano, E., Vestergaard, C.L.: Randomized reference models for temporal networks. arXiv preprint arXiv:1806.04032 (2018)
9. Gutiérrez-Gómez, L., Delvenne, J.-C.: Multi-hop assortativities for network classification. J. Complex Netw. **7**(4), 603–622 (2019)
10. Harrigan, M., Fretter, C.: The unreasonable effectiveness of address clustering. In: 2016 International IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 368–373. IEEE (2016)
11. Interdonato, R., Atzmueller, M., Gaito, S., Kanawati, R., Largeron, C., Sala, A.: Feature-rich networks: going beyond complex network topologies. Appl. Netw. Sci. **4**(1), 1–13 (2019)
12. Joseph, S.M., Citraro, S., Morini, V., Rossetti, G., Stella, M.: Cognitive network science quantifies feelings expressed in suicide letters and reddit mental health communities. arXiv preprint arXiv:2110.15269 (2021)
13. Jourdan, M., Blandin, S., Wynter, L., Deshpande, P.: Characterizing entities in the bitcoin blockchain. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 55–62. IEEE (2018)
14. Kondor, D., Pósfai, M., Csabai, I., Vattay, G.: Do the rich get richer? An empirical analysis of the bitcoin transaction network. PLoS ONE **9**(2), e86197 (2014)
15. Kovanen, L., Kaski, K., Kertész, J., Saramäki, J.: Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc. Natl. Acad. Sci. **110**(45), 18070–18075 (2013)
16. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. Soc. Netw. Anal. Min. **8**(1), 1–29 (2018)
17. Lee, E., Karimi, F., Wagner, C., Jo, H.-H., Strohmaier, M., Galesic, M.: Homophily and minority-group size explain perception biases in social networks. Nat. Hum. Behav. **3**(10), 1078–1087 (2019)
18. 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), e0136497 (2015)
19. McPherson, M., Smith-Lovin, L., Cook, J.M.: Homophily in social networks. Annual review of sociology, Birds of a feather (2001)
20. Molloy, M., Reed, B., Newman, M., Barabási, A.-L., Watts, D.J.: A critical point for random graphs with a given degree sequence. In: The Structure and Dynamics of Networks, pp. 240–258. Princeton University Press (2011)
21. Moody, J.: Race, school integration, and friendship segregation in America. Am. J. Sociol. **107**(3), 679–716 (2001)
22. Morini, V., Pollacci, L., Rossetti, G.: Toward a standard approach for echo chamber detection: reddit case study. Appl. Sci. **11**(12), 5390 (2021)
23. Newman, M.E.J.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003)
24. Parmentier, P., Viard, T., Renoust, B., Baffier, J.-F.: Introducing multilayer stream graphs and layer centralities. In: International

Conference on Complex Networks and Their Applications, pp. 684–696. Springer (2019)
25. Peel, L., Delvenne, J.-C., Lambiotte, R.: Multiscale mixing patterns in networks. Proc. Natl. Acad. Sci. **115**(16), 4057–4062 (2018)
26. Pelechrinis, K., Wei, D.: VA-index: quantifying assortativity patterns in networks with multidimensional nodal attributes. PLoS ONE **11**(1), e0146188 (2016)
27. Posfai, M., Barabási, A.-L.: Network Science. Cambridge University Press, Cambridge (2016)
28. Rabbany, R., Eswaran, D., Dubrawski, A.W., Faloutsos, C.: Beyond assortativity: proclivity index for attributed networks (PRONE). In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer (2017)
29. Rathore, A.S., Mutalikdesai, M.R., Patil, S.: Analyzing trust-based mixing patterns in signed networks. In: International Conference on Asian Digital Libraries, pp. 63–72. Springer (2013)
30. Remy, C., Rym, B., Matthieu, L.: Tracking bitcoin users activity using community detection on a network of weak signals. In: International Conference on Complex Networks and Their Applications, pp. 166–177. Springer (2017)
31. Rossetti, G., Citraro, S., Milli, L.: Conformity: a path-aware homophily measure for node-attributed networks. IEEE Intell. Syst. **36**, 25–34 (2021)
32. Sapiezynski, P., Stopczynski, A., Lassen, D.D., Lehmann, S.: Interaction data from the Copenhagen networks study. Sci. Data **6**(1), 1–10 (2019)
33. Sepulvado, B., Wood, M.L., Fridmanski, E., Wang, C., Chandler, M.J., Lizardo, O., Hachen, D.: Predicting homophily and social network connectivity from dyadic behavioral similarity trajectory clusters. Soc. Sci. Comput. Rev. 0894439320923123 (2020)
34. Shrum, W., Cheek Jr, N.H., MacD, S.: Friendship in school: gender and racial homophily. Sociol. Educ. 227–239 (1988)
35. Simard, F.: On computing distances and latencies in link streams. In: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 394–397. IEEE (2019)
36. Simard, F.: Evaluating metrics in link streams. Soc. Netw. Anal. Min. **11**(1), 1–16 (2021)
37. Simard, F., Magnien, C., Latapy, M.: Computing betweenness centrality in link streams. arXiv preprint arXiv:2102.06543 (2021)
38. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., et al.: High-resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE **6**(8), e23176 (2011)
39. Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B., Comte, B., Voirin, N.: Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PLoS ONE **8**(9), e73970 (2013)
40. Zhou, B., Xin, L., Holme, P.: Universal evolution patterns of degree assortativity in social networks. Soc. Netw. **63**, 47–55 (2020)