Attributed Stream-Hypernetwork Analysis: Homophilic Behaviors in Pairwise and Group Political Discussions on Reddit

Andrea Failla[1], Salvatore Citraro[1(✉)], and Giulio Rossetti[2]

[1] Computer Science department, University of Pisa, Pisa, Italy
a.failla@studenti.unipi.it, salvatore.citraro@phd.unipi.it
[2] KDD Laboratory, ISTI, National Research Council, Pisa, Italy
giulio.rossetti@isti.cnr.it

Abstract. Complex networks are solid models to describe human behavior. However, most analyses employing them are bounded to observations made on dyadic connectivity, whereas complex human dynamics involve higher-order relations as well. In the last few years, hypergraph models are rising as promising tools to better understand the behavior of social groups. Yet even such higher-order representations ignore the importance of the rich attributes carried by the nodes. In this work we introduce ASH, an Attributed Stream-Hypernetwork framework to model higher-order temporal networks with attributes on nodes. We leverage ASH to study pairwise and group political discussions on the well-known Reddit platform. Our analysis unveils different patterns while looking at either a pairwise or a higher-order structure for the same phenomena. In particular, we find out that Reddit users tend to surround themselves by like-minded peers with respect to their political leaning when online discussions are proxied by pairwise interactions; conversely, such a tendency significantly decreases when considering nodes embedded in higher-order contexts - that often describe heterophilic discussions.

Keywords: High-order networks · Dynamic networks · Feature-rich networks · Data analysis

1 Introduction

Several human behaviors can be described through the lens of complex network analysis. From homophily [20] to the time-evolving nature of social groups [14], complex networks can shed lights on the laws governing nodes' wiring patterns and their evolution as time goes by as well. However, the intrinsic nature of network connectivity can not explicitly go beyond dyadic patterns, i.e., coupling between pairs of nodes. Such pairwise constraints must be taken into account when aiming to investigate the behavior of social groups and their complex

dynamics. Hypernetwork science is the new, cutting-edge line of research that aims to address higher-order structures for representing and analyzing complex systems [1,3,27]. In this work we consider such higher-order structures involved in dynamic and attribute-aware environments as well. This way we propose to study a phenomenon along three main dimensions: topology, dynamic features and node attributes, plus any preferred combination of these. To this purpose we introduce *ASH*, an *A*ttributed *S*tream-*H*ypernetwork model to represent higher-order temporal networks with attributive information on nodes. To test ASH's potentiality, we infer the attribute-rich higher-order temporal structure of group political discussions on the well-known Reddit platform. In particular, we focus on data collected from the debate between Trump supporters and anti-Trump citizens as described in [18] along three main topics: gun legalization, minorities discrimination and general debates on the US political sphere. We aim to explore to what extent users are homogeneously embedded in group discussions when modeling the higher-order structure, where *homogeneity* is computed as the fraction of users within the higher-order neighborhood of a target node that share the same attribute value of this latter one. Then, we aim to compare these results to the ones obtained by mining the pairwise network only.

The rest of the work is organized as follows. Section 2 sums up the principal literature on the three main complex network contexts surrounding this work, namely dynamic, node-attributed, and higher-order networks. Section 3 introduces our Attributed Stream-Hypernetwork model. Section 4 discusses our main results on the selected datasets. Finally, Sect. 5 concludes the work and discusses promising lines of research left open for the future.

2   Related Work

We provide in the following an overview of the main enriched network models used in this work. First, we discuss dynamic and node-attributed network representations [16]; then, we sum up the emerging contributions about higher-order representations for complex systems.

Dynamics of networks. The temporal dimension is often available from network data representing human behaviors. However, choosing a proper representation for modeling interaction dynamics is not a straightforward task. Friendships are persistent over time, whereas face-to-face interactions involve a certain duration, and e-mails or financial transactions are even instantaneous. Different temporal semantics impose different modeling strategies [14,24]. We can identify dynamic network models by selection of different criteria: (i) stability, e.g., the snapshot sequence from a time-window aggregation strategy [6,23]; (ii) duration, e.g., the interval graph model [14]; and (iii) immediacy, e.g., the stream graph [17], where a dynamic network comes as a stream of both temporal nodes and interactions. Stream graphs, in particular, can easily extend and generalize classic centrality measures on static graphs [26], and multi-layer structure as well [21]. Among the most interesting and cutting-edge tasks on dynamic networks we can mention community detection [24], link prediction [13], and mixing pattern estimation [10].

Networks with metadata. Similarly to temporal information, metadata or attributes describing nodes' properties are often available from network data. Metadata-enriched models can support augmented mining and analyses about the relationships between the structural and the attributive information in complex systems. Most of the works focusing on the interplay between structure and attributes aim to study the correlation between the two dimensions, searching for bridges between tabular and networked data [28]. Node attributes can be fruitfully used for improving the community detection task, where both tight connectivity and label-homogeneity within communities need to be guaranteed [8]. They can be also leveraged for estimating heterogeneous mixing patterns in complex networks [22,25]. The distribution of metadata surrounding a single node, e.g., the distribution of features within the node's ego-network, is particularly useful in the node classification and in the link prediction tasks [4].

Higher-order networks. Although traditional network science mostly addressed analyses on pairwise interactions, many network dynamics can be better modelled by higher-order representations involving relations between groups of nodes. As an emerging line of research [3,27], the expressive power of such representations is yet largely unexplored. The interest in the physics of higher-order interactions is growing [2], and it has been mainly explored in the context of diffusion analysis, e.g., studying social contagion with simplicial complexes [15], in time-varying settings as well [7]. Higher-order structures varying in time are an important and emerging trend of research [5]. In [12], three distinct types of hypergraph neighborhoods are introduced: star, radial and contracted ego-networks. The star ego-network of a node u is the set of all the hyperedges that include u, whereas the other two include also the hyperedges whose nodes within are all u's neighbors (radial) or further extends u's neighborhood to the hyperedges that interact with nodes outside of u's ego-network (contracted). The differences between the three ego-networks is interesting when studying them from a temporal perspective: some hyperedges included in the radial and contracted neighborhoods may have begun a time before the target node u joined the context [12]. In most of such analyses, the higher-order structure of static/dynamic networks is addressed by investigating datasets originally designed for graph-based analysis, thus one of the most intriguing future challenges is the inference of statistically significant higher-order interactions from complex systems [19]. Moreover, higher-order-based techniques are emerging to generalize well-known graph-based techniques or to conservatively shift to them, as in the case of s-line graph analysis for hypergraph models [1].

3   Attributed Stream-Hypergraphs

To study dynamic high-order social interactions, simply borrowing results from the existing literature is not enough. Hypergraphs and simplicial complexes, to name the most used high-order representation frameworks [3,27], have both strengths and weaknesses. None of them has been adequately defined in the

presence of evolving topologies. Indeed, their applicability to online social environments needs to be carefully analyzed to understand if the constraint they come with aligns with the semantics expressed by social interaction networks. Moreover, individuals embedded in a social system can often be characterized by multiple features—*profiles* that contextualize some of the key properties playing a role for social interactions (e.g., nationality, gender, age...). To start filling the existing gap in high-order dynamic and feature-rich modeling of social systems, here we propose the framework of Attributed Stream Hypergraphs.

Definition 1. *[Attributed Stream Hypergraph (ASH)]* Let $\mathcal{S} = (T, V, W, E, L)$ be a stream hypergraph, where:

– $T = [A, \Omega]$ is the set of discrete time instants, with A and Ω the initial and final instants;
– V is the set of the nodes of the temporally flattened hypergraph;
– $W \subseteq T \times V$ is the set of temporal nodes;
– $E \subseteq T \times V^n$ is the set of temporal hyperedges such that $(t, N) \in E$ implies that $N \subseteq V$ and $\forall u_i \in N, (t, u_i) \in W$;
– L is the set of temporal node attributes such that $L(t, u)$ with $(t, u) \in W$ and $t \in T$, identifies the set of categorical values associated to u at time t.

ASHs are a conservative extension of well known modeling frameworks, namely hypergraphs [3] and stream graphs [17]. For instance, temporal nodes and temporal edges are a peculiarity of the stream graph representation, where the nature of nodes and edges is analyzed with respect to the time they appear in the temporal stream. A node is a temporal entity that can be present or absent at a certain time in the stream, so that a node is said to be *equal to 1*—i.e., it is represented as a whole quantity as in a classic static graph—only if it is present all the time in the stream. Similarly, an edge accounts for 1 only if it is present all the time in the stream. From both hypergraphs and stream graphs ASH inherits several analytical peculiarities and, from their union, it is able to provide novel insights that the original models are not able to unveil independently. Temporal hyperedges indeed are an example of such novelty, where the temporal presence of an interaction is accounted for higher-order groups and not only pairwise interactions. Moreover, integrating time evolving node attributes, it allows to study not only how individuals' characteristics change (e.g., opinions, political leaning) but also how such changes relate/affect the topological structure surrounding them.

In the following we provide a first example of how ASH can be used to study real world phenomena. Our analysis will focus on some aspects of the three dimensions modeled by ASH: high-order topology, node semantics and time.

4   Experiments

Sociopolitical Reddit data. We focus on data collected from the debate between Trump supporters and anti-Trump citizens during the first two and

half years of Donald Trump's presidency, covering a period between January 2017 and July 2019. Data collection, users' ideology inference and network construction are properly described in the reference paper [18] (thus identifying three users' families: protrump, antitrump and neutral). Leveraging the original temporal network,[1] we infer the hypergraph structure by means of all the maximal cliques. As in the original analysis about this dataset, [18], we consider a time-window of six months when analyzing system interactions' dynamics. Average statistics for the pairwise graphs are shown in Table 1.

The three main topics considered are the followings:

- *Gun Control*: this topic is identified by collecting lists of subreddits that either support gun legalization or are against it;
- *Minorities Discrimination*: identified by considering groups that promote gender/racial/sexual equality and groups showing more conservative attitudes;
- *Political Sphere*: identified by covering different US political ideologies such as Republicans, Democrats, Liberals, and Populists.

Table 1. Network statistics (averaged across semesters). size of the network in terms of nodes and edges, number of users with a pro-trump, anti-trump or neutral leaning score.

Topic	# Nodes	# Edges	# Pro-trump	# Anti-trump	# Neutral
Gun Control	4991	15,298	3346	1645	–
Minorities Discrimination	5540	12,605	3318	2222	–
Political Sphere	4509	7079	1280	2395	834

Among these topics, the first two ones are controversial/polarizing sociopolitical issues, whereas the latter one is intended to offer a more broad spectrum of the US political debate.

Analytical setting. We set a four-fold framework to analyze ideology homogeneity from different perspectives as in the following:

i. we promote an analysis on dyadic interactions, and we measure how much users are homogeneously embedded in their pairwise ego-networks;
ii. we shift the focus from individual users to groups, and we measure the homophily of such groups/contexts modelled as hyperedges;
iii. we return to individual users, adopting the single user point of view while measuring how much an user is embedded in the contexts/hyperedges where he/she participates;
iv. we introduce a time-aware analysis to track stability or variations in ideology homogeneity.

---

[1] Original network data available at https://github.com/virgiiim/EC_Reddit_Case_Study.

As a preliminary question we aim to explore whether different behaviours emerge among individual users (i) and groups (ii): can higher-order interactions capture patterns which are invisible to dyadic interactions? Then, we aim to understand the role of single users in the several contexts where they participate (iii), as a meeting point between the two previous points: can higher-order neighborhoods capture different patterns that graph ego-networks cannot? Finally, the focus on interactions' dynamics (iv) would allow us to track stable or mutable patterns as time goes by.

It should be noted that computations in (i) and (iii) are different from (ii). In (i) and (iii) we use a measure of homogeneity to estimate target nodes' similarity within nodes own contexts. The focus is on single nodes and we aim to measure the homogeneity of the contexts with respect to the political leaning of a specific target node. In detail, the homogeneity within the context of a target node with respect to the node's own ideology is computed as the relative frequency of the attribute value carried by the node within the context that embeds it:

$$Homogeneity(u) = \frac{|\{v|v \in c \wedge l_u = l_v\}|}{|c|}, \tag{1}$$

where u is the target node with l_u as attribute value, c is the context where u is embedded, and l_v is the attribute value of a node v that belongs to c. In the pairwise network, c is the first-order neighborhood, namely the set of nodes adjacent to u. In the hypergraph, c is the set of hyperedges of u's star ego-network. Since a hyperedge includes several nodes, we must associate to the context expressed by the hyperedge a characteristic value of such a context; coherently with the choice to measure the homogeneity with respect to a target node neighborhood, we set as the characteristic value of a hyperedge the most frequent attribute value of the nodes participating in it.

Finally, the homogeneity of the hypergraph is defined as the average of all its nodes' homogeneities:

$$Homogeneity = \frac{1}{|V|} \sum_{u \in V} Homogeneity(u). \tag{2}$$

Conversely, in (ii) the focus is on hyperedges' homogeneities. Thus, we use a particular instantiation of the homogeneity measure, namely the purity [11], that computes the relative frequency of the most frequent attribute value within the hyperedge:

$$Purity(e) = \frac{\max_{l \in L}(\sum_{v \in e} l_v)}{|e|}, \tag{3}$$

where e is the target hyperedge, and v is a node participating in it with l_v as attribute value. This way we only capture the characteristic behaviour of a target hyperedge.

Finally, the purity of the hypergraph is defined as the average of all its hyperedges' purities:

$$Purity = \frac{1}{|E|} \sum_{e \in E} Purity(e). \tag{4}$$

Results

Pairwise ego-networks reveal both homophilic and heterophilic users' preferences. Fig. 1 outlines graph ego-networks' homogeneities in the three topics considered. Results are aggregated over the semesters. The analysis on pairwise interactions captures both homophilic and heterophilic patterns. In *GunControl*, there emerges heterogeneity in antitrump and protrump users, where the latter ones appear to be more likely to engage with similar users. Conversely, an overall strong homophilic distribution emerges in *Minority*, whose homogeneity values for both antitrump and protrump nodes mostly gather around the upper bound. A strong difference is visible in *Politics*, where protrump and neutral users are less homogeneous than antitrump ones. All these observations are coherent with the analyses performed on the original data paper [18], where in *Minority* and *Politics* it is more likely to observe echo-chambers—oriented towards a protrump political leaning in *Minority*, and antitrump in *Politics*, while *GunControl* discussions are less polarized.
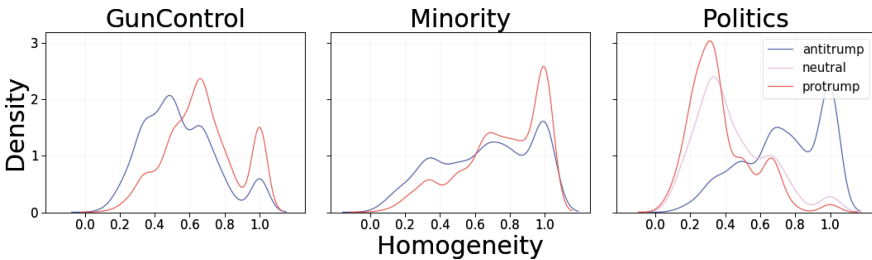

Fig. 1. KDE distributions of pairwise ego-networks' homogeneity among the three different Reddit communities.

Hyperedges' purity emphasizes homogeneous discussions. Figure 2(a) gives a global overview of ideology homogeneity within the contexts/hyperedges in the three topics considered. Results are aggregated over the semesters. The bimodal distributions observed in all the categories reflect similarly to graph ego-networks what already emerged on the same data while identifying echo chambers at the meso-scale community level [18]: *GunControl does not present strongly polarized communities [i.e., echo chambers] among different semesters* (p. 12) [18] as well as it seems that only a bunch of contexts present quite perfect purity (Fig. 2(a), leftmost); *[in Minority] on average, more*
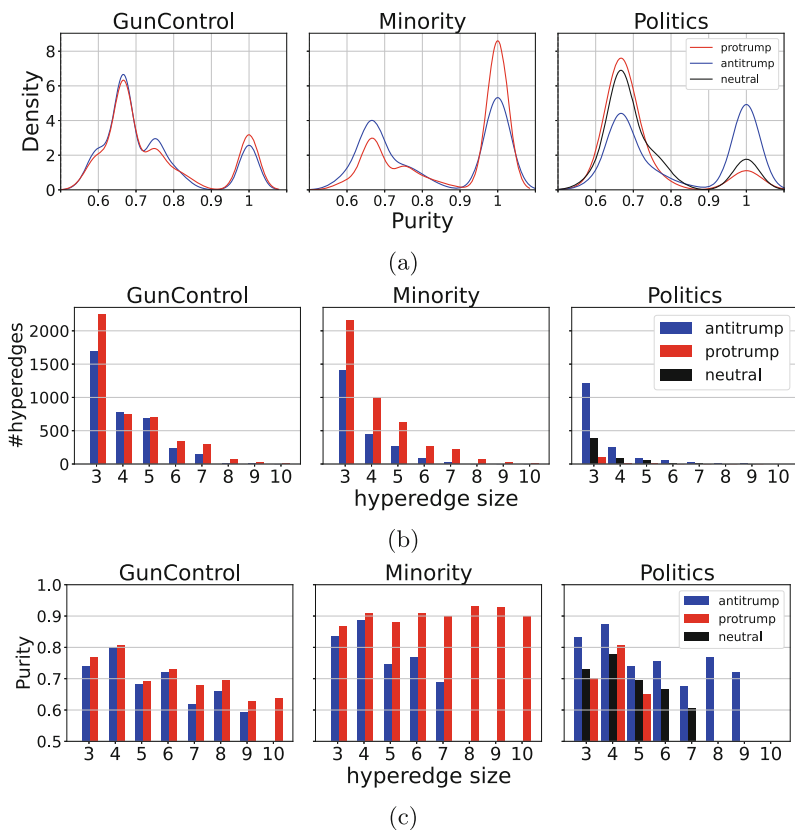
Fig. 2. KDE distributions of hyperedges' purity (a), number of pure hyperedges (b) and average purity (c) in function of hyperedge size among the three different Reddit communities.

than half of total users are trapped in echo chambers (p. 12) [18], and hyperedge purities show a quite similar pattern as well, with a tendency of protrump users to form more homogeneous groups (Fig. 2 (a), center); also *Politics* presents high homogeneity contexts, where antitrump users are more likely to form homogeneous groups (Fig. 2(a), rightmost). Nonetheless, this analysis leveraging higher-order contexts reveals more heterogeneity in the dataset and emphasizes different behaviours among users, even when focusing on the purity only. Moreover, all these patterns are observed with respect to the hyperedge size as well. Figure 2(b, c) highlight, respectively, the number (b) and the average purity (c) of pure-groups in function of the group size. For instance, in *Minority* we observe that only protrump pure discussions involve groups with more than 7 participants, and that they are quite pure, 0.9. The same does not happen in *GunControl*, while in *Politics* the biggest contexts involve antitrump users only but with a lower purity than the one of protrump users in *Minority*.

Users are involved in heterogeneous debates.
As can be observed in Fig. 3, the topics show diversified behaviours when the
analysis shifts to star egos, which describe the hypernetworks at the meso-scale
level while maintaining local information about individual nodes. Indeed, there
is no more trace of the bimodal patterns observed in Fig. 2(a), with the sole
exceptions of antitrump nodes in *Minority* and neutral ones in *Politics*. The
key insight, however, relates to the heterogeneity of user debates. While engag-
ing in relatively homogeneous contexts (Fig. 2), it seems that users find them-
selves in rather mixed collections of debates. That is to say, although homophilic
behaviour is highlighted in most debates (i.e., hyperedges), the set of contexts a
node is involved in (i.e., its star) is generally diversified w.r.t. ideology/political
leaning. This is especially true in *GunControl*, where protrump users appear
to engage in a more heterogeneous set of debates than their counterparts, as
opposed to what was noted in Fig. 2(a). The same holds for *Politics*, which dis-
plays a peak in heterogeneous protrump stars while the antitrump ones show
more homophilic behavior. *Minority*, instead, still shows strong homogeneity
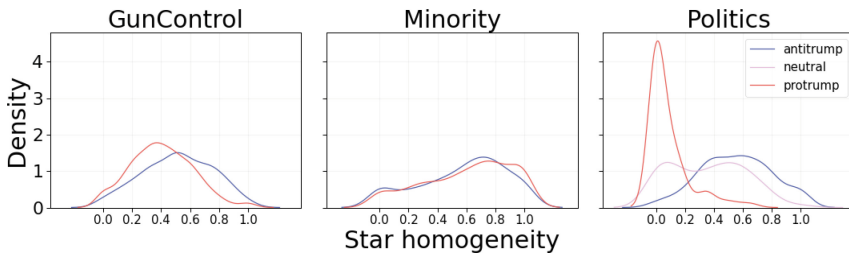traits for both antitrump and protrump users, thus confirming previous obser-
vations.

Fig. 3. KDE distributions of hypergraph star ego-networks' homogeneity among the
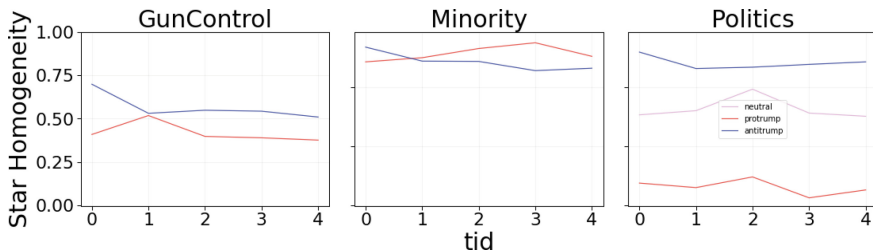three different Reddit communities.

Fig. 4. Average hypergraph star ego-networks' homogeneity over time among the three
different Reddit communities.

Interactions' dynamics: users' preferences tend to be consistent in time.
As far as the temporal dimension is concerned, a certain degree of consistency
w.r.t. debates homogeneity/heterogeneity can be observed. As a matter of fact,
the average star homogeneity outlines almost-flat trends (Fig. 4), indicating
minor variations. Here, *GunControl* and *Minority* reveal near-constant hetero-
geneity/homogeneity for both political alignments; lastly, *Politics* displays only
a small bump during the third semester concerning protrump and neutral users.

5   Discussion and Conclusions

In this work we described and applied ASH, our new framework to represent
streams of higher-order interactions enriched with node attributes as Attributed
Stream-Hypernetworks. This way a complex phenomenon can be described along
several dimensions, namely topology, node-features, and interaction dynamics,
plus any combination of these as preferred by the analyst of a specific domain
[16]. We applied our framework to study group political discussions on Reddit,
analyzing to what extent politically oriented users are homogeneously embedded
in group discussions with respect to group's representative political leaning. Even
if not mainly related to the identification/analysis of echo chambers, this work
gave us another perspective to interpret the homophilic patterns emerging from
controversial political debates on social platforms. Our results can support the
claim of users aggregation in similar/homophilic contexts/hyperedges, but this
result may not necessarily involve segregation [9]. In fact, while applying our
framework to the Reddit communities, we observed strong homophilic behav-
iors among groups/hyperedges with respect to users' political leaning. However,
while focusing on the preferences of single nodes, namely on how much a target
node is homogeneously embedded with respect to the representative political
leaning of the groups/hyperedges it belongs to, we mostly observe a relevant
decrease in nodes' homophilic behaviors. As a consequence, we observe that
users prefer participating in contexts whose representative leaning is different
than the target node's own label, although hyperedges are strongly homophilic
per se. Interestingly, this pattern can not be observed when looking at the pair-
wise ego-networks only. Finally, regarding the interaction dynamics, we mainly
observe stability and consistency over time with respect to users' political lean-
ing. In future works, we plan to work with more datasets that can better involve
changes along the temporal dimension. Also, we plan to define domain-specific
measures based on a solid characterization of node profiles, leveraging them to
better analyze higher-order structure dynamics. Lastly, we plan to formalize
more consistently the ASH framework, as already done for recent generaliza-
tions of dynamic networks as stream graphs [17], and we plan to focus on the
constraints that stream hypergraphs could eventually raise, such as the issues of
under/over fitting social data or the robustness of the measures to missing data.

References

1. Aksoy, S.G., Joslyn, C., Marrero, C.O., Praggastis, B., Purvine, E.: Hypernetwork science via high-order hypergraph walks. EPJ Data Sci. **9**(1), 16 (2020)
2. Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., et al.: The physics of higher-order interactions in complex systems. Nature Phys. **17**(10), 1093–1098 (2021)
3. Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.G., Petri, G.: Networks beyond pairwise interactions: structure and dynamics. Phys. Rep. **874**, 1–92 (2020)
4. Bhagat, S., Cormode, G., Muthukrishnan, S.: Node classification in social networks. In: Social Network Data Analytics, pp. 115–148. Springer (2011)
5. Cencetti, G., Battiston, F., Lepri, B., Karsai, M.: Temporal properties of higher-order interactions in social networks. Sci. Rep. **11**(1), 1–10 (2021)
6. Chiappori, A., Cazabet, R.: Quantitative evaluation of snapshot graphs for the analysis of temporal networks. In: International Conference on Complex Networks and Their Applications, pp. 566–577. Springer (2021)
7. Chowdhary, S., Kumar, A., Cencetti, G., Iacopini, I., Battiston, F.: Simplicial contagion in temporal higher-order networks. J. Phys. Complex. **2**(3), 035019 (2021)
8. Chunaev, P.: Community detection in node-attributed social networks: a survey. Comput. Sci. Rev. **37**, 100286 (2020)
9. Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proc. Nat. Acad. Sci. **118**(9) (2021)
10. Citraro, S., Milli, L., Cazabet, R., Rossetti, G.: $\{\backslash Delta\}$-conformity: multi-scale node assortativity in feature-rich stream graphs (2021). arXiv:2111.15534
11. Citraro, S., Rossetti, G.: Identifying and exploiting homogeneous communities in labeled networks. Appl. Netw. Sci. **5**(1), 1–20 (2020)
12. Comrie, C., Kleinberg, J.: Hypergraph ego-networks and their temporal evolution. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 91–100. IEEE (2021)
13. Divakaran, A., Mohan, A.: Temporal link prediction: a survey. New Gen. Comput. **38**(1), 213–258 (2020)
14. Holme, P., Saramäki, J.: Temporal networks. Phys. Rep. **519**(3), 97–125 (2012)
15. Iacopini, I., Petri, G., Barrat, A., Latora, V.: Simplicial models of social contagion. Nature Commun. **10**(1), 1–9 (2019)
16. Interdonato, R., Atzmueller, M., Gaito, S., Kanawati, R., Largeron, C., Sala, A.: Feature-rich networks: going beyond complex network topologies. App. Netw. Sci. **4**(1), 1–13 (2019)
17. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. Social Netw. Anal. Mining **8**(1), 1–29 (2018)
18. Morini, V., Pollacci, L., Rossetti, G.: Toward a standard approach for echo chamber detection: reddit case study. Appl. Sci. **11**(12), 5390 (2021)

19. Musciotto, F., Battiston, F., Mantegna, R.N.: Detecting informative higher-order interactions in statistically validated hypergraphs. Commun. Phys. **4**(1), 1–9 (2021)
20. Newman, M.E.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003)
21. 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)
22. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. Proc. Nat. Acad. Sci. **115**(16), 4057–4062 (2018)
23. Ribeiro, B., Perra, N., Baronchelli, A.: Quantifying the effect of temporal resolution on time-varying networks. Sci. Rep. **3**(1), 1–5 (2013)
24. Rossetti, G., Cazabet, R.: Community discovery in dynamic networks: a survey. ACM Comput. Surv. (CSUR) **51**(2), 1–37 (2018)
25. Rossetti, G., Citraro, S., Milli, L.: Conformity: a path-aware homophily measure for node-attributed networks. IEEE Intel. Syst. **36**(1), 25–34 (2021)
26. Simard, F., Magnien, C., Latapy, M.: Computing betweenness centrality in link streams (2021). arXiv:2102.06543
27. Torres, L., Blevins, A.S., Bassett, D., Eliassi-Rad, T.: The why, how, and when of representations for complex systems. SIAM Rev. **63**(3), 435–485 (2021)
28. Zanin, M., Papo, D., Sousa, P.A., Menasalvas, E., Nicchi, A., Kubik, E., Boccaletti, S.: Combining complex networks and data mining: why and how. Phys. Rep. **635**, 1–44 (2016)