Emerging Models and Paradigms in Network Science

Part #3: Modelling Interactions over Time: Stream Graphs and Link Streams

Ph.D. Course at University of Calabria

27–30 May 2024

Francesco Cauteruccio

DIEM, University of Salerno

fcauteruccio@unisa.it

francescocauteruccio.info

Introduction

- Graph theory, social network analysis, and network science all share a common concept: the usage of a *language* for describing networks.
- This language consists of elementary yet powerful concepts:
  - Node, node degree, paths, density, cliques, etc.
- It forms the basis of network studies, and different dialects have been designed to deal with other contexts:
  - The dialect of time (*temporal networks*), or the dialect of layered interactions (*multilayer networks*).

Interactions over Time I

- Interactions that take place over time are ubiquitous:
  - Contacts, shopping, travels, traffic, etc.
  - These are streams of nodes and links active during specific periods of time.
- As we have seen with temporal networks, the most common approach is to model them
  - *by sequences of graphs* → **data is aggregated**,
  - *by labeled graphs* → **data is available through edges**.
- These methods introduce new forms of complexity to deal with.

Interactions over Time II

- Recently, a different approach has emerged:
  - Instead of transforming interactions into graphs, we transform graph theory into a theory of interactions over time.
  - Such an approach allows one to deal directly with interactions over time, similar to what graph theory does for relations.
- This approach culminated in the definition of *link streams* and *stream graphs*:
  - These are **not** graphs or networks,
    - But when no dynamics are involved (e.g., no time), the concept is equivalent to a classical graph.
  - Both elementary and complex graph concepts are defined for them, too!
    - Relations between graph properties still hold for stream properties.

Definitions I

- A (simple undirected) graph $G = (V, E)$ is defined by a finite set of nodes V and a set of links $E \subseteq V \otimes V$. $uv \in E$ means that $u, v \in V$ are linked together in G.
  - Here, $X \otimes Y$ denotes the set of unordered pairs of distinct elements of X and Y, while $X \times Y$ is the set of ordered pairs.
- A (simple undirected) *stream graph* is defined as $S = (T, V, W, E)$, where
  - V is a finite set of nodes,
  - T is a measurable set of time instants,
  - $W \subseteq T \times V$ is a set of temporal nodes,
  - $E \subseteq T \times V \otimes V$ is a set of links, such that $(t, uv) \in E$ implies $(t, u) \in W$ and $(t, v) \in W$.
    - $(t, v) \in W$ means that node v is present at time t in S, and $(t, uv) \in E$ means that nodes u and v are linked together at time t in S.

Definitions II

- Some notation shortcuts:
  - $v_t = 1$ if $(t, v) \in W$, and $v_t = 0$ otherwise;
    - If $v_t = 1$, we say that *node v is present in S at time t*.
  - $uv_t = 1$ if $(t, uv) \in E$, and $uv_t = 0$ otherwise.
    - If $uv_t = 1$, we say that *nodes u and v are linked together at time t in S*.
  - T_v is the set of time instants at which v is present, T_{uv} is the set of time instants at which uv is present.
    - $T_{uv} \subseteq T_u \cap T_v$.
  - V_t is the set of nodes present at time t, E_t is the set of links present at time t.
- If nodes are present all the time, i.e., $T_v = T, \forall v \in V$, then S is a *link stream*, and is denoted with $L = (T, V, E)$ (with $W = T \times V$ implicitly).
  - There is no dynamics on nodes, therefore S is fully defined by T, V, E.

Examples - Stream Graph

- $S = (T, V, W, E)$
  - $T = [0, 10] \subseteq \mathbb{R}$
  - $W = [0, 10] \times \{a\} \cup ([0, 4] \cup [5, 10]) \times \{b\} \cup [4, 9] \times \{c\} \cup [1, 3] \times \{d\}$
  - $E = ([1, 3] \cup [7, 8]) \times \{ab\} \cup [4.5, 7.5] \times \{ac\} \cup [6, 9] \times \{bc\} \cup [2, 3] \times \{bd\}$
  - The dotted lines represent the time a node is active:
    - $T_a = [0, 10]$, $T_b = [0, 4] \cup [5, 10]$, etc.
    - $T_{ab} = [1, 3] \cup [7, 8]$, $T_{ac} = [4.5, 7.5]$, etc.

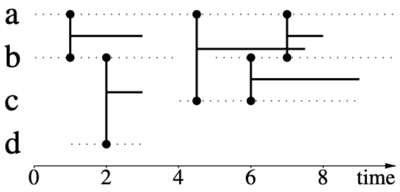

Figure: A stream graph. Source: [2]

Examples - Link Stream

- $L = (T, V, E)$
  - $E = ([0,4] \cup [6,9]) \times \{ab\} \cup [2,5] \times \{ac\} \cup [1,8] \times \{bc\} \cup [7,10] \times \{bd\} \cup [6,9] \times \{cd\}$
  - Nodes are active all the time:
    - $T_{ab} = [0,4] \cup [6,9]$, $T_{ac} = [2,5]$, etc.

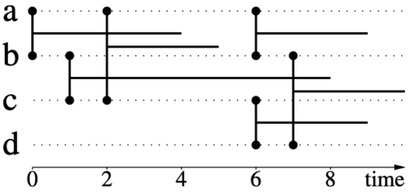

Figure: A link stream. Source: [2]

Induced Graphs

- Given a stream graph $S = (T, V, W, E)$, we define $G_t = (V_t, E_t)$ as the graph induced by S at time t.
- Also, $G(S) = (\{v \mid T_v \neq \varnothing\}, \{uv \mid T_{uv} \neq \varnothing\}) = (\bigcup_{t \in T} V_t, \bigcup_{t \in T} E_t)$ is the *graph induced by* S.
  - Its nodes are those present in S and they are linked together in $G(S)$ if there exists a time instant t in T such that they are linked together in S.

Measures and Properties

- Stream graphs represent a key formalism for studying jointly the dynamics and structure of interactions.
- To do so, we need measures to characterize their properties, such as:
  - Coverage, size, duration, uniformity, and compactness;
  - Density;
  - Topological structures and properties:
    - Neighborhood, degree, cliques, clustering, paths, and distances;
    - Connectedness;
  - Centralities.

Coverage

- In a stream graph $S = (T, V, W, E)$, nodes may not be present all the time in general.
  - Therefore, W may differ significantly from $T \times V$.
- To capture this aspect, we define the *coverage of S* as

$$cov(S) = \frac{|W|}{|T \times V|}$$

  - For instance, the previous example of stream graph S has $cov(S) = \frac{26}{40} = 0.65$.
- Note that $cov(S) = 1$ if and only if all nodes are present all the time, and therefore, S is a link stream.
- If all links are present all the time, then there is no significant distinction between S and $G(S)$, and we say S is a *graph-equivalent stream*.

Size and Duration I

- Given S, we define the *contribution of node v* as $n_v = \frac{|T_v|}{|T|}$.
- Then, the *number of nodes n in S* is defined as

$$n = \sum_{v \in V} n_v = \frac{|W|}{|T|}$$

  - v in V accounts for 1 node only if it is present in S all the time.
- Similarly, we define the *contribution of a pair of nodes uv* as $m_{uv} = \frac{|T_{uv}|}{|T|}$, and the *number of links m in S* as

$$m = \sum_{uv \in V \otimes V} m_{uv} = \frac{|E|}{|T|}$$

  - uv in $V \otimes V$ accounts for 1 link only if it is present in S all the time.

Size and Duration II

- The node and link contributions of a time instant t are $k_t = \frac{|V_t|}{|V|}$ and $l_t = \frac{|E_t|}{|V \otimes V|}$, respectively.
- Then, the node duration k in S and the link duration l in S are

$$k = \frac{|W|}{|V|} \quad \text{and} \quad l = \frac{|E|}{|V \otimes V|}$$

  - Note that n is the expected value of $|V|$ when we take a random time $t \in T$;
  - Likewise, m, k, and l are the expected values of $|E_t|$, $|T_v|$, and $|T_{uv}|$ when we take a random time $t \in T$, a random node $v \in V$, or a random pair of nodes $uv \in V \otimes V$.

Uniformity and Compactness I

- The presence at the same time of pairs of nodes can be better highlighted with the concept of uniformity.
- Given S, we define the uniformity of S as

$$\mathbb{U}(S) = \sum_{uv \in V \otimes V} \frac{|T_u \cap T_v|}{|T_u \cup T_v|}$$

  - If $\mathbb{U}(S) = 1$, then for all nodes $u, v \in V$, $T_u = T_v$.
- Uniformity can also be defined for any pair of nodes $u, v \in V$ as the Jaccard coefficient

$$\mathbb{U}(u, v) = \frac{|T_u \cap T_v|}{|T_u \cup T_v|}$$

  - It measures the overlap between the presence times of u and v.

Uniformity and Compactness II

- Given S, we define $S' = (T', V', W, E)$ such that

$$T' = [\min\{t \mid \exists(t, v) \in W\}, \max\{t \mid \exists(t, v) \in W\}]$$

and

$$V' = \{v \mid \exists(t, v) \in W\}$$

we define the *compactness of S* as

$$c(S) = \frac{|W|}{|T' \times V'|} = cov(S')$$

  - If $c(S) = 1$, then we say S is compact, that is, the presence times of all nodes are in the same interval of T.

Density I

- In graphs, the density indicates the probability that when one takes a random pair of nodes, there is a link connecting them.
- In stream graphs, we define the density as the probability that when one takes a random time and two random nodes, such that a link may exist between them at this time, the link indeed exists.
  - Pragmatically speaking, it is the fraction of possible links that do exist.

$$\delta(S) = \frac{\sum_{uv \in V \otimes V} |T_{uv}|}{\sum_{uv \in V \otimes V} |T_u \cap T_v|}$$

Density II

- Density can also be defined for *a pair of nodes* $uv \in V \otimes V$, for *a node* $v \in V$, and *at a time instant* $t \in T$ as follows:

$$\delta(uv) = \frac{|T_{uv}|}{|T_u \cap T_v|}$$

  - i.e., the probability that there is a link when u and v are both present.

$$\delta(v) = \frac{\sum_{u \in V, u \neq v} |T_{uv}|}{\sum_{u \in V, u \neq v} |T_u \cap T_v|}$$

  - i.e., the probability that a link between v and any other node exists whenever this is possible.

$$\delta(t) = \frac{|E_t|}{|V_t \otimes V_t|}$$

  - i.e., the probability that a link exists between any two nodes present at time t.

- All these measures are equal to 0 by definition if the denominator is 0.

Sub-streams and Clusters

- Let $S = (T, V, W, E)$ and $S' = (T', V', W', E')$ be two stream graphs.
- Their *intersection* is the stream graph $S \cap S' = (T \cap T', V \cap V', W \cap W', E \cap E')$
  - It is their largest (with respect to inclusion) common sub-stream.
- Their *union* is the stream graph $S \cup S' = (T \cup T', V \cup V', W \cup W', E \cup E')$
  - It is the smallest stream graph having both S and S' as sub-streams.
- A cluster C of $S = (T, V, W, E)$ is a subset of W.
  - The set of links between nodes involved in C are $E(C) = \{(t, uv) \in E : (t, u) \in C \wedge (t, v) \in C\}$.
  - We denote by $S(C) = (T, V, C, E(C))$ the *sub-stream of S induced by* C.

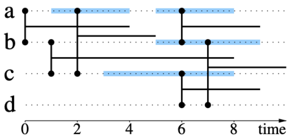

Figure: In blue the cluster $C = ([1,4] \cup [5,8]) \times \{a\} \cup [5,9] \times \{b\} \cup [3,8] \times \{c\}$

Cliques

- A *clique* of a stream graph S is a cluster C of S with density 1.
  - All pairs of nodes involved in C are linked in S whenever both are involved in C.
- A clique C is maximal if there is no other clique C' such that $C \subset C'$.
- A clique is compact if its induced sub-stream is compact.
- A clique is uniform if its induced sub-stream is uniform.

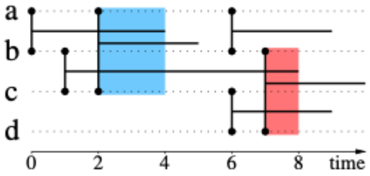

Figure: Two maximal compact cliques involving three nodes of the link stream. Source: [2]

Neighborhood and Degree

- In a graph, the neighborhood $N(v)$ of a node v is the cluster $N(v) = \{u : uv \in E\}$, and the degree $d(v)$ of v is $|N(v)|$.
- In a stream graph $S = (T, V, W, E)$, we define the *neighborhood of a node v* as

$$N(v) = \{(t, u) : (t, uv) \in E\}$$

and the *degree* $d(v)$ *of* v as the number of nodes in $N(v)$.

  - Equivalently, $d(v) = \frac{|N(v)|}{|T|} = \sum_{u \in V} m_{uv}$.
  - This means that each node u contributes to the degree of v proportionally to the duration of its links with v.

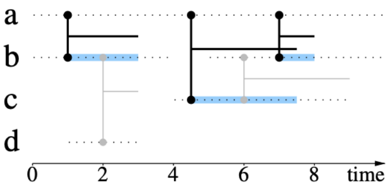

Figure: Neighborhoods and degrees of node. $N(a) = ([1, 3] \cup [7, 8]) \times \{b\} \cup [4.5, 7.5] \times \{c\}$ is in blue, leading to $d(a) = \frac{3}{10} + 310 = 0.6$

Neighborhood and Degree I

- The *average node degree of S* is

$$d(V) = \frac{1}{n} \cdot \sum_{v \in V} n_v \cdot d(v) = \sum_{v \in V} \frac{|T_v|}{|W|} \cdot d(v)$$

  - Here, the contribution of each node v to the average node degree of S is weighted by its presence duration $|T_v|$.

- These definitions generalize graph concepts to stream graphs; however, the temporal features can be considered in other generalizations.

Neighborhood and Degree II

- Given S, we define:
  - The *instantaneous neighborhood of a node v at time t* as $N_t(v) = \{u : (t, uv) \in E\}$,
  - The *instantaneous degree of v at time t* as the number of nodes in $N_t(v)$.
  - If v is not involved in S at time t, then $N_t(v) = \emptyset$ and $d_t(v) = 0$.
- We can average instantaneous degrees in two ways:
  - Over all nodes in V, thus $d(t) = \sum_{v \in V} \frac{d_t(v)}{|V|}$, and we call this the degree at t;
  - Over nodes in V_t only, thus $d(t) = \sum_v \frac{1}{|V_t|} d_t(v)$, and we call this the expected degree at time t (which is exactly the average degree of G_t).
- For averages over all V and T, we obtain

$$d(S) = \sum_{v \in V} \frac{1}{|V|} d(v)$$

which is the average instantaneous degree of v at time t for random $(t, v) \in T \times V$, and is called the degree of S.

Clustering Coefficient and Transitivity Ratio I

- Given S, we define the *clustering coefficient of a given node v* as the density of its neighborhood

$$cc(v) = \delta(N(v)) = \frac{\sum_{uw \in V \otimes V} |T_{vu} \cap T_{vw} \cap T_{uw}|}{\sum_{uw \in V \otimes V} |T_{vu} \cap T_{vw}|}$$

  - Pragmatically speaking, $cc(v)$ is the probability when one takes two random neighbors u and w of v at time t, i.e., a random $(t, uw) \in T \times V \otimes V$, such that (t, vu) and (t, vw) are in E, that u is linked to $w \in S$ at time t, i.e., that $(t, uw) \in E$.

- Then, the *node clustering coefficient of S* is the average clustering coefficient of all its nodes, weighted by their presence in S, that is

$$cc(S) = \frac{1}{n} \sum_{v \in V} n_v \cdot cc(v) = \sum_{v \in V} \frac{|T_v|}{|W|} \cdot cc(v)$$

Clustering Coefficient and Transitivity Ratio II

- The transitivity ratio $tr(S)$ of S is the probability, when one takes a random connected triplet, that it is a triangle

$$tr(S) = \frac{\triangledown}{\vee}$$

  - $\vee$ is the set of all connected triplets of S, and $\triangledown$ is the set of all triangles in S.
  - We say that $(t, (u, v, w)) \in T \times (V \times V \times V)$, $u \neq v \neq w$ is a connected triplet if at time t there is both a link between u and v and between v and w, i.e., $(t, uv) \in E$, and $(t, vw) \in E$. If in addition, there is a link between u and w, then it is a triangle.

k-cores

- The k-core of a graph $G = (V, E)$ is its largest cluster $C^k \subseteq V$ such that for all $v \in C^k$, $d(v) \geq k$ in the subgraph $G(C^k)$ of G induced by C^k.
  - This cluster is unique for a given k and $C^{k+1} \subseteq C^k$ for all k.
  - One can compute it by iteratively removing from G all elements of V of degree lower than k.
  - The 0-core of G is V, and the k-core contains all cliques of size $k + 1$ of G.
- The k-core of a stream graph $S = (T, V, W, E)$ is its largest cluster $C^k \subseteq W$ such that for all $(t, v) \in C^k$, $d_t(v) \geq k$ in the sub-stream $S(C^k)$ of S induced by C^k.
  - One can compute it by iteratively removing from S all elements of W of instantaneous degree lower than k.
  - The 0-core of S is W, and the k-core contains all compact cliques of S involving $k + 1$ nodes.

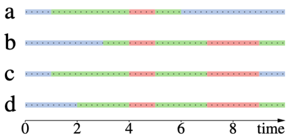

Figure: The 2-core of L is highlighted in red. L refers to the example given previously. Source: [2]

Paths and Distances I

- In a stream graph $S = (T, V, W, E)$, a *path* P from $(\alpha, u) \in W$ to $(\omega, v) \in W$ is a sequence

$$(t_0, u_0, v_0), (t_1, u_1, v_1), \dots, (t_k, u_k, v_k)$$

  of elements of $T \times V \times V$ such that:
  - $u_0 = u, v_k = v$,
  - $t_0 \geq \alpha, t_k \leq \omega$,
  - 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, $[t_i, t_{i+1}] \times \{v_i\} \subseteq W$.
- We say that P involves $(t_0, u), (t_k, v)$, and (t, v_i) for all $i \in [1, k-1]$ and $t \in [t_i, t_{i+1}]$.
- Also, we say that P *starts at* t_0, *arrives at* t_k, *has length* $k+1$, and *duration* $t_k - t_0$.

Paths and Distances II

- The notion of reachability is a bit different from graphs:
  - If there exists a path from (α, u) to (ω, v) in S, we say that (ω, v) is *reachable* from (α, u), and we denote it as $(\alpha, u) \dashrightarrow (\omega, v)$.
  - We say that a node v is reachable from u if there exists α and ω such that $(\alpha, u) \dashrightarrow (\omega, v)$, and we denote it as $u \dashrightarrow v$.
  - Reachability is not symmetric nor transitive!
- Paths in stream graphs are quite different from paths in graphs:
  - Their temporal nature makes them not symmetric.
    - The existence of a path from u to v does not imply the existence of a path from v to u.
  - They have a length but also a duration, thus the cost of reaching a node from another one must be defined differently.

Paths and Distances III

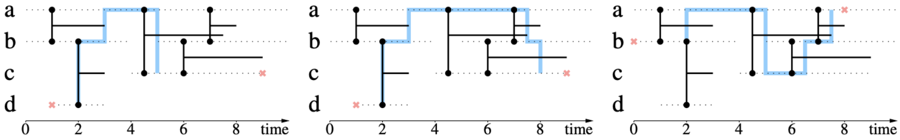

Fig. 13 Paths in a stream graph. Left: a path P_1 from $(1, d)$ to $(9, c)$: $P_1 = (2, d, b), (3, b, a), (5, a, c)$. This path has length 3 and duration 3. Center: another path P_2 from $(1, d)$ to $(9, c)$: $P_2 = (2, d, b), (3, b, a), (7.5, a, b), (8, b, c)$. This path has length 4 and duration 6. Right: a path P_3 from $(0, b)$ to $(8, a)$: $P_3 = (2, b, a), (5, a, c), (6.5, c, b), (7.5, b, a)$. This path has length 4 and duration 5.5

Figure: Source: [2]

Shortest and Fastest Paths I

- We say that P is a *shortest path* from (α, u) to (ω, v) if it has minimal length, and we call this length the *distance* from (α, u) to (ω, v), denoted by $((\alpha, u), (\omega, v))$.
  - The distance (u, v) from u to v is the minimal such distance for all α and ω in T, and a shortest path from u to v is a path from u to v with length (u, v).
- We say that P is a *fastest path* from (α, u) to (ω, v) if it has minimal duration, and we call this duration the *latency* from (α, u) to (ω, v), denoted by $((\alpha, u), (\omega, v))$.
  - The latency (u, v) from u to v is the minimal such latency for all α and ω, and a fastest path from u to v is a path from u to v with duration (u, v).

Shortest and Fastest Paths II

- We denote by $\mathcal{T}_\alpha(u,(t,v)) = \omega - \alpha$ the *time to reach* (t,v) from u at time α, where $\omega \leq t$ is the minimal value such that there is a path from (α,u) to (ω,v) in S and $[\omega,t] \subseteq T_v$. We call such a path a *foremost path* from (α,u) to (t,v).
- Shortest paths are optimal regarding the number of hops, while fastest paths are optimal regarding the duration between starting and arrival time.
  - This captures the following intuition: if someone (at a given time) wants to go to another city by train (and arrive before a given time), this person may want to minimize the number of train changes (shortest path), the total time they spend traveling (fastest path), or the time at which they will arrive at the destination (foremost path).
- Several extensions can be defined for paths [2]:
  - For instance, γ-paths consider paths with a maximum cost in terms of duration of γ.

Δ-analysis I

- In some contexts, directly studying the stream graph induced by a dataset makes little sense.
- Imagine we are representing a set of phone calls or instant messages:
  - In the former case, nodes generally have only zero or one link at a time, thus leading to an instantaneous degree of 0 or 1.
  - In the latter, links are instantaneous, leading to a density equal to 0.
- In such cases, we are interested in the fact that nodes interact regularly, typically at least once every Δ units of time, for a given Δ.
  - For instance, two individuals call each other at least once a day, two sensors detect each other at least once a day, etc.
- To do so, the usual approach consists in using Δ as a parameter to define notions to describe the data.

Δ-analysis II

- Examples:
  - n_Δ and m_Δ can be defined as the expected number of nodes and links, respectively, present in a randomly chosen time interval of duration Δ in T.
  - The Δ-degree $d_\Delta(v)$ of a node v is the expected number of neighbors during a randomly chosen time interval of duration Δ in T.
- Stream graphs are useful to provide a more general way to do so.
- Given a stream graph $S = (T, V, W, E)$ with $T = [\alpha, \omega]$ and a value $\Delta \leq \omega - \alpha$, we define $S_\Delta = (T_\Delta, V, W_\Delta, E_\Delta)$ as a stream graph such that:
  - $T_\Delta = [\alpha + \frac{\Delta}{2}, \omega - \frac{\Delta}{2}]$,
  - $W_\Delta = \{(t', v) : t' \in T_\Delta, \exists (t, v) \in W \text{ such that } |t' - t| \leq \frac{\Delta}{2}\}$, and
  - $E_\Delta = \{(t', uv) : t' \in T_\Delta, \exists (t, uv) \in E \text{ such that } |t' - t| \leq \frac{\Delta}{2}\}$
- In other words, a node is present at time t' in S_Δ whenever it is present in S at a time t in $[t' - \frac{\Delta}{2}, t' + \frac{\Delta}{2}]$, i.e. $T_{\Delta_v} = T_\Delta \cap \{t' : \exists t \in T_v, |t' - t| \leq \frac{\Delta}{2}\}$.

Δ-analysis III

- The properties of S_Δ actually are equivalent to the Δ-properties of S [2]:
    - Therefore, one may conduct Δ analysis of S by transforming it into S_Δ.

Conclusion

- Stream graphs represent a formalism to deal directly with the intrinsically temporal and structural nature of interactions over time.
  - It becomes natural to describe how elementary metrics evolve over time.
- Data that would benefit from being represented through this formalism are countless:
  - Mobility traces, financial transactions, etc.
  - At the time of writing, more than 240 papers cited [2] and numerous application contexts have been investigated through the lens of stream graphs, as well as different extensions have been proposed [1].
- Stream graphs lie at the crossroad of two very rich scientific areas: graph theory and time series analysis.
  - Time series concepts could be generalized to stream graphs.

Exercise

- Let's try playing with stream graphs
- The exercises should be done in Python using iPython Notebook.
- Three steps:
  1. Define a `StreamGraph` class
  2. Make an instance of such class.
  3. Create a function that given a stream graph and one of its nodes, computes the neighborhood and the degree of this node.
- A notebook to start with is available at `exercises/stream-graph-0.ipynb`

Thanks for your attention!

Francesco Cauteruccio, Ph.D.
University of Salerno

fcauteruccio@unisa.it
https://francescocauteruccio.info

References I

Andrea Failla, Salvatore Citraro, and Giulio Rossetti.
Attributed stream hypergraphs: temporal modeling of node-attributed high-order interactions.
Applied Network Science, 8(1):31, 2023.

Matthieu Latapy, Tiphaine Viard, and Clémence Magnien.
Stream graphs and link streams for the modeling of interactions over time.
Social Network Analysis and Mining, 8:1–29, 2018.