Supplementary Material for A Simple Embedding for Classifying Networks with a few Graphlets

Luce le Gorrec
Department of Mathematics and Statistics
University of Strathclyde
Glasgow G1 1XH, United Kingdom
luce.le-gorrec@strath.ac.uk

Philip A. Knight
Department of Mathematics and Statistics
University of Strathclyde
Glasgow G1 1XH, United Kingdom
p.a.knight@strath.ac.uk

I. Introduction

This document provides additional background to [1]. It contains a thorough description of the network benchmarks used in its numerical experiments and we provide an identification of all the k-node graphlets. We expound the theoretical connection between the paper's feature selection procedure and Principal Component Analysis (PCA). Finally, we provide details and justifications for the parameter choices used in the numerical experiments in [1].

II. The Network Benchmark

The experiments in [1] use a network dataset which we have curated. In brief, the dataset is populated with unweighted directed static networks without self-loops all of which have more than 50 nodes. They are real world networks from 4 different application fields, namely food webs, electronic circuits, discourse structures and social networks. Here we give information about the networks, the repositories where we found them as well as any post-processing we have applied to them to obtain networks that are well suited to our analysis.

A. Food webs.

A food web represents the consumer–resource relations between species in an ecosystem. That is, there exists a relation between species a and b if individuals from species a consume individuals from species b—most often, this means a eats b but other kinds of relations may exist, for instance, if species a is a parasite of the host species b.

In a network that represents a food web, the nodes are the species, and the edges represent the consumer–resource relations. We consider directed edges such that there is an edge from a to b if species a consumes species b. Networks from this field may be weighted, with weights representing either a level of confidence in the consumer–resource relation, or the quantity of species b that is consumed by species a. When meeting such weighted networks, we have kept all edges but removed the weight. Self-loops occasionally occur, in cannibal species for instance. These are excised. In some food webs the prey and predators are partitioned into two disjoint groups of species. We have discarded such networks as the bipartite structure is the dominant characteristic.

In total we have collected 70 food webs from the following repositories.

- 33 food webs have been extracted from the website GlobalWeb [2]. We have filtered all the food webs to keep only the non-bipartite networks with more than 50 nodes. Among this first set, we have only kept the networks in which we, with limited biological knowledge, could interpret the relations—that is, numerical values between species (where relevant), as well as which species were prey, and which were predators.
- 21 food webs have been extracted from [3]. This article analyses different kinds of food web to understand the impact of taking into account parasitism. It first observes food webs without parasites and considers free-living species only (species that do not need a host). Then the study adds parasitic species, but without concomitant links (a concomitant link is a link that indicates that if predator a eats preys b, it also eats parasites hosted by b). Finally, food webs with parasitic species and concomitant links are analysed. The study analyses seven different ecosystems, and for each ecosystem, three food webs are provided.
- We have used the data from [4] in a similar way to [2], to generate 4 food webs. In the first we keep the links corresponding to predation only, the second keeps host–parasite relations, the third includes concomitant predation relations and a final network includes all the relations.
- 11 food webs have been extracted from Pajek [5]. In the Pajek dataset, edges are weighted, and some nodes are not formal species—for instance, in each network, there are three nodes labelled Input, Output and Respiration. We have decided to keep all these nodes when we extract

This project was supported by the Royal Academy of Engineering and the Office of the Chief Science Advisor for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme.

the networks, and to take into account all the edges, whatever their weight.
- The last food web in our benchmark comes from the Konnect repository [6], taken from the study [7].

The food webs from [2], [4] were found using the Icon Colorado Repository [8].

B. Electronic Circuits.

Here we consider an electronic circuit as an object composed of one or more inputs that provide one or more outputs by passing through logic gates (AND, NOR, etc.). An electronic circuit can be represented by a network where a node may be an input, an output or a logic gate, and an edge indicates that a signal enters or leaves a logic gate. For instance, the relation

$$\Sigma_1 = AND(\Sigma_2, \Sigma_3)$$

will be represented by one node, with two edges entering the node, representing signals Σ_2 and Σ_3, and one edge leaving the node that represents signal Σ_1. An example for the electronic circuit s27 from [9] is provided in Figure 1.

For this field, 52 networks with more than 50 nodes have been extracted from two classical benchmarks, namely ISCAS'89 [9] and ITC'99 [10], found in repositories [11] and [8].

C. Discourse Structures.

Segmented Discourse Representation Theory [12] is used to represent the rhetorical relations between discourse units by means of networks. An example illustrating how such a network is built, extracted from [13] and [14], is shown in Figure 2. The discourse units can be of two kinds: elementary discourse units (EDUs), or complex discourse units (CDUs), corresponding to discourse units composed of several EDUs.

For this field, we have extracted the networks from the STAC dataset[1] [15], which is a corpus of chat conversations between players from an online version of the board game "The Settlers of Catan". Two corpora are available to download: a linguistic one, in which there are only human being exchanges (players and observers), and a situational one, which is the linguistic corpus completed with information about the game given by the computer by means of canned messages[2]. For both corpora, we have extracted all the networks with more than 50 nodes, which has resulted in a dataset of 195 networks—22 from the linguistic corpus, 173 from the situational corpus. As shown in Figure 2, the edges are labelled. We do not consider these labels in our study.

D. Social relations.

The term "social relation" covers a wide range of relation types, from friendship relations on online social networks to physical contacts; and from commercial exchanges to genealogical relations, or even domination among a group of animals. In our study, we have limited ourselves to human relations that are directed and can be symmetrical—this excludes, for instance, genealogical or student–teacher relationships. The resulting dataset consisting of 81 networks that can be roughly divided into two kinds of relations, namely feelings and exchanges.

Networks representing feelings are mainly based on surveys of groups of individuals, such as students from the same school or co-workers, who were asked to name their friends, or who they turn to to ask for advice or support, etc. In such networks, the edges may be weighted according to the strength of the feeling. These weights may sometimes be negative if the individuals can expressed both positive and negative feelings—trust and distrust, friend or foe. In these cases we only take into account the existence of a feeling, and we draw an edge whenever the individual a expresses a feeling—weak or strong, positive or negative—for individual b. The edges may also be labelled according to the nature of the relation if several kinds of relations have been tested. In this case, we have created one network per relation type, and one network merging all the relation types.

Networks representing exchanges (emails, phone calls, messages on online social websites) may be dynamic (the times of exchanges are known), may have multi-edges (several exchanges from individual a to individual b are possible). Self-loops are sometimes allowed since, for example, one can send an email to oneself. We have transformed these networks into static unweighted networks by adding an edge from a to b so long as there is at least one exchange at some time from a to b. Finally, we remove self-loops.

The social networks we have used come from many sources (including surveys) as we detail below.

- 29 networks come from the study [16], which is a survey held in several schools. Students were asked to name their closest friends. For each friend named, the student was asked whether he/she participated in any of five activities with the friend. In the resulting network, nodes are students, and there is an edge from a to b if student a named b as a close friend. The edges are labelled by the activities that a claims to have done with b. We limited ourselves to the studies of the first 30 schools, omitting school 3 which surveyed fewer than 50 students. In post-processing we removed edge labels. Our source for data is the repository [17].
- 2 networks come from the study [18], where students in a high school were asked to name their friends. Two surveys took place, one in 1957 and one in 1958. Source data can be found in [17].
- 1 network from the study [19]. This is a survey based on the same principles as above: students living in a residence are asked to name their friends within the residence, giving a strength to their relationship. The initial network is available from [17].
- 9 networks come from the study [20], a survey of Dutch secondary schools. Pupils have been asked to name other pupils with whom they share an emotional support

[1] https://www.irit.fr/STAC/corpus.html

[2] Such as "It's <player id> to roll the dice", "<player id> ends its turn", etc.

l1: # 4 inputs
l2: # 1 outputs
l3: # 3 D-type flip-flops
l4: # 3 inverters
l5: # 8 gates(1 ANDs +1 NANDs +2 ORs +4 NORs)
l6: INPUT(G0)
l7: INPUT(G1)
l8: INPUT(G2)
l9: INPUT(G3)
l10: OUTPUT(G17)
l11: G5 = DFF(G10)
l12: G6 = DFF(G11)
l13: G7 = DFF(G13)
l14: G14 = NOT(G0)
l15: G17 = NOT(G11)
l16: G8 = AND(G14,G6)
l17: G15 = OR(G12,G8)
l18: G16 = OR(G3,G8)
l19: G9 = NAND(G15,G16)
l20: G10 = NOR(G14,G11)
l21: G11 = NOR(G5,G9)
l22: G12 = NOR(G1,G7)
l23: G13 = NOR(G2,G12)

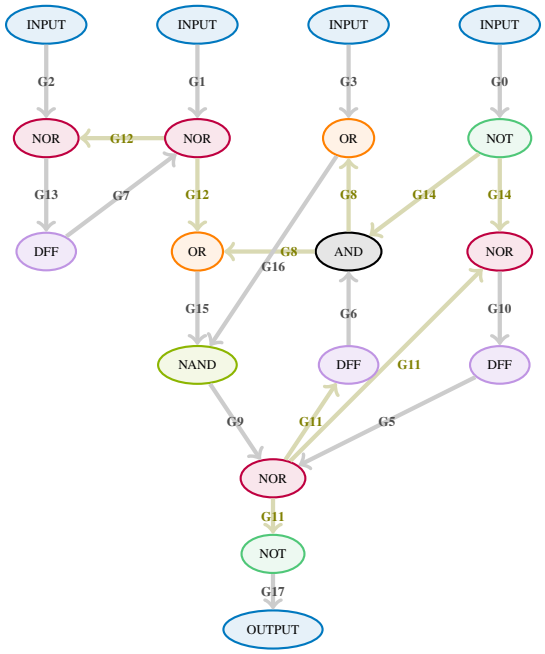

Fig. 1. Left: Netlist of the electronic circuit s27 from [9]. Right: The resulting network, with labels displayed. Each gate/input/output corresponds to one node. The edges are signal transfers. For instance, signal G14, which is the result of input signal G0 going through a NOT gate (*l14*), enters both a NOR gate with signal G11 to create signal G10 (*l20*), and an AND gate with signal G6 to create signal G8 (*l16*).

γ_1 : John had a lovely evening last night.
γ_2 : He had a great meal.
γ_3 : He ate salmon.
γ_4 : He devoured lots of cheese.
γ_5 : He won a dancing competition.
Γ_6 : $\gamma_2 \cup \gamma_5$
Γ_7 : $\gamma_3 \cup \gamma_4$

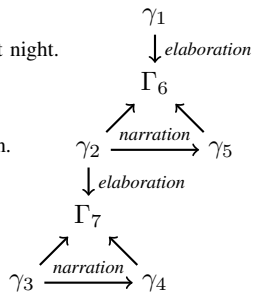

Fig. 2. Left: A short text divided into discourse units. The EDUs are identified by a γ, the CDUs by a Γ. Right: Discourse structure of the text with a network representation. The edge labels correspond to rhetorical relations. An edge with no label links an EDU to the CDU it is a member of.

relationship (either as giver or receiver). 36 classrooms were surveyed, resulting in 36 networks. We kept the 9 classrooms whose largest weak component contained at least 50 nodes. These networks are available at [21].

- 6 networks come from a survey by Vladis Krebs of the IT department of a Fortune 500 company. Workers were asked to name other workers with whom they have certain kinds of relationships (co-workers on multiple projects, sought for advice for decision making, sought for technical expertise, discussion of customers issues). Their relationships were assigned a value ("for each kind of [relation], each individual reported the frequency of contact with each of the other individuals (Yearly or less, Quarterly, Monthly, Weekly, Daily or more"). In the resulting networks (one per relationship type plus one for the merger of all relationships), nodes are workers, and there is an edge from a to b if b has been named by a for this type of relationship. These sources were taken from [17]. Note that for customer issues, the answers from the last individual have been only partially reported (11 answers from 55). We have kept the network with the incomplete answers.
- 4 networks come from a similar study to Krebs but in a law firm [22], with questions about advice, co-workers and friendship. The relations are not valued. Once again, we have one network per relation kind plus one that merges all the relations. These networks were found at [17].
- 4 networks come from another work place survey, this time of GPs in Illinois [23]. The questions asked were: "When you need information or advice about questions of therapy where do you usually turn?", "Who are the three or four physicians with whom you most often find yourself discussing cases or therapy in the course of an ordinary week – last week for instance", "Would you tell me the first names of your three friends whom you see most often socially?". This resulted in 4 networks (advice, discussion, friendship, and the merger of all relationships). These sources were found at [17].
- 2 networks come from a survey of a research team in a manufacturing company [24]. Given a list of other coworkers, the researchers were asked to answer two questions. The first question was "Please indicate the extent to which the people listed below provide you with information you use to accomplish your work" on the

scale : "0: I Do Not Know This Person/I Have Never Met this Person; 1: Very Infrequently; 2: Infrequently; 3: Somewhat Infrequently; 4: Somewhat Frequently; 5: Frequently; and 6: Very Frequently". The second question asked the researchers to rate the following assertion "I understand this person's knowledge and skills. This does not necessarily mean that I have these skills or am knowledgeable in these domains but that I understand what skills this person has and domains they are knowledgeable in" on the scale: "0: I Do Not Know This Person/I Have Never Met this Person; 1: Strongly Disagree; 2: Disagree; 3: Somewhat Disagree; 4: Somewhat Agree; 5: Agree; and 6: Strongly Agree." We have built two networks from this survey, one per question, where a node is a researcher and there is an edge from a to b if for researcher b, a has provided an answer greater or equal to 3. Initial data were found at [25].

- 1 network comes from a survey of drug users in Hartford, Connecticut [26]. Nodes are drug users, and directed edges represent acquaintance. Data can be found at [27].
- 1 network comes from a survey of prison inmates who were asked to name their closest friends within the prison [28]. In the resulting network, nodes are prison inmates and there is an edge from a to b if b is designated as a close friend by a. The network can been found at [5].
- 1 network comes from a survey that lists the frequent visits among families living in a same neighbourhood [29], [30]. The visits can be of 3 kinds, namely "visiting relation as ordinary", "visits among kin", and "visits among ritual kin". In the network, nodes are the families and there is an edge from a to b if family a frequently visits family b. A post-processing is applied to remove the labels on the edges (that is, we keep only the information about frequent visits, and do not take into account the type of visits). Initial data can be found at [5].
- 2 networks come from another survey [29] based on the same principles as above. One of the networks was built in exactly the same way (but on another neighbourhood), and has been treated in the same way. For the other, each family was asked which other families they would notify in case of a family member's death. In this network, each node is a family, and there is an edge from a to b if family a answered they would advise family b in case of death. Initial data were taken from [5].

We also have social networks extracted from websites in which edges represent feelings.

- 1 network, presented in [31], is an online social network called Anybeat, where people can choose to follow other people. Nodes are users, edges denote the fact a user follows another user. Initial data came from [32].
- 3 networks come from the technology website Slashdot-Zoo, where users can mark other users as friend or foe. The three networks correspond to three different time-stamps. One has been presented in [33] and can be found at [6]. The two others have been presented in [34] and can be found at [35]. In these networks, nodes are users, and an edge from a to b indicates that user a has labelled user b as his/her friend or his/her foe, resulting in labelled edges. We have post-processed the networks to remove these labels, so that we only keep the information that user a has expressed a feeling for user b.
- 2 networks, discussed in [36] and available at [35], focus on the hyperlinks between communities (called subreddits) of the social news website Reddit. A subreddit can make a post that mentions a post from another subreddit. Two networks are built, one for the mention of another post in the title of the current post, the other if the mention is in the post body. The mention can be negative, neutral or positive. In the networks, a node is a subreddit, and there is an edge from a to b if a post from subreddit a mentions a post from subreddit b in its title (respectively its body). Post-processing was required for these networks. There are weights (-1 if the post from a is negative toward b) and timestamps (time where the post appeared) associated with the edges in these networks, that we choose to ignore.
- 1 network, presented in [37] and available at [6], is a Who-trusts-whom social network taken from a consumer review website. Nodes are users and an edge from a to b means that user a has indicated he/she trusts user b.
- 1 network, presented in [38], is based on an online community platform for developers of free software. Users can give certification to other users or to themselves corresponding to the level of trust a user has to another user (or themselves). Three levels of certification are possible. In this network, nodes are users, edges are trust relations. In post-processing, self-loops and edge weights corresponding to certification level are removed. The network before post-processing can be found at [6].
- 2 networks, presented in [39], [40], are two Who-trusts-whom networks of traders extracted from two Bitcoin platforms. Users of the platforms can rate other members with integers from -10 (total distrust) to +10 (total trust). Post-processing was required for these networks. We chose to ignore the weights and time-stamps associated with the edges in these networks. That is, if member a rates a grade of $k \in \{\pm 1, ..., \pm 10\}$ to user b at time t, in the resulting network we only see one static edge from a to b. The networks (before post-processing) have been extracted from [35].

Finally, we also have some social networks that represent exchanges between individuals. These exchanges are mainly emails or phone calls.

- 1 network, presented in [41], is a subset of the posts left by Facebook users on their friends' walls. Nodes are users, an edge from a to b indicates user a posts something on the wall of user b. Post-processing has been used to remove edge weights (a user may post several times on the same wall), and self-loops (posting on one's own wall). The initial network has been found at [6].

- 1 network, presented in [42], is the list of email exchanges among members of a European research institution. Nodes are email addresses and there is an edge from a to b if a has sent at least one email to b. This network come from [35].
- 1 network, presented in [43], is based on the same principle as the network above, but in a mid-sized manufacturing company instead of a European research institution. Post-processing was applied because the initial network was dynamic: times when emails were sent were known. We create an edge from a to b if a sent at least one email to b. Original data are available at [6].
- 2 networks, presented in [44], represent the communication network of the Linux Kernel mailing list. Nodes are email addresses, there is an edges from a to b if a had made a reply to a request of b. A post-processing has been applied since the initial network is dynamic, with the time of the reply provided. We create a simple unweighted edge from a to b if at least one reply has been made from a to a request from b. The resulting network has 3 connected components with more than 50 nodes, but the smallest one is a simple path (and hence, is bipartite). We have kept the two largest only. The original network can be found at [6].
- 1 network, presented in [45] and available at [6], documents messages sent on a social website among students from the University of California, Irvine. Nodes are students, and there is an edge from a to b if a sent a message to b. Again, post-processing has been applied since the initial network was dynamic: times when messages were sent were recorded. We create an edge from a to b if a sent at least one message to b.
- 2 networks, analysed in [46], are the results of an investigation of university students within the Copenhagen Networks Study. Two networks have been generated. One corresponds to phone calls, the other ones to text messages. Nodes are students and there is an edge from a to b if a has called/texted b. The networks are dynamic (times when messages or phone calls have been passed is known), and weighted (number of text messages, number of phone calls plus length of phone calls). We have post-processed these networks to make them unweighted. We did not take into account missed calls.
- 1 network, presented in [47], is the result of an investigation of hashish and cocaine importers who were identified as members of the so-called Caviar network. Nodes represent people actively implicated in the Caviar network who were caught in a police surveillance net, and there is an edge from a to b if communications from a to b have been observed during the investigation. A post-processing has been applied because the initial network is valued by the strength of contact between two individuals. We have kept all edges, whatever the strength of the contact, and made them unweighted. The original dataset is available at [48].

# nodes	mixing param. μ	exp. of power law for community size: t_2	average in-deg. $\overline{k}$	max. in-deg. k_{max}
$n \times 1000$, $n = 1, \ldots, 5$	0.6	-1	5	100
	0.6	-2	5	100
	0.8	-1	5	100
	0.8	-1	5	50
	0.8	-1	10	50
	0.8	-1	10	100

TABLE I
TABLE OF PARAMETERS USED TO GENERATE RANDOM GRAPHS WITH LFR GENERATOR. GIVEN A SET OF PARAMETERS $(n, \mu, t_2, \overline{k}, k_{max})$, TWO RANDOM NETWORKS HAVE BEEN GENERATED.

In total we obtained 81 networks for this database. As for the food webs, the listing provided by [8] was very helpful for sourcing the data.

E. Random Networks

In Section VI of [1] a class of 60 randomly generated networks were used. These were constructed using the so-called LFR graph generator from [49][3]. This generator was initially designed to test community detection algorithms by generating random modular networks with features observed in real-world networks (notably heterogeneous distributions of node degrees and community sizes described by power laws). Several parameters need to be set up and we have built the networks with a mixture of default and bespoke value choices. We summarise these parameters in Table I. Further details can be found in [49].

III. GRAPHLET IDENTIFIER

Studies that focus on graphlets need to identify these graphlets with a unique identifier. We have chosen one of the several ways to identify graphlets that exists in the literature. First we give the general formula to derive an identifier for a graph, then we explain how it is applied to the graphlets.

Let $G = (V, E)$ represent an unweighted directed graph such that $V = \{v_1, ..., v_k\}$ is the set of nodes ($|V| = k$) and E is the set of edges where (v_i, v_j) indicate there exists an edge from v_i to v_j in G. Let $\mathbf{A} \in \{0, 1\}^{k \times k}$ be the adjacency matrix of G,

$$a_{i,j} = 1 \iff (v_i, v_j) \in E.$$

Let $\mathbf{B} \in \mathbb{R}^{k \times k}$ be a matrix whose entry (i, j) is given by:

$$b_{i,j} = 2^{(i-1)\times k + (j-1)},$$

Then, the graph G can be identified by the integer

$$id_G = \mathbf{e}^T (\mathbf{B} \circ \mathbf{A}) \mathbf{e}, \tag{III.1}$$

with $\circ$ the Hadamard product, or pairwise matrix multiplication, and $\mathbf{e}$ a vector of ones.

When considering graphlets, we presume two isomorphic subgraphs correspond to a same graphlet. In other words, in a graphlet, node label is immaterial. Hence, to derive a graphlet

[3]Using the "Package 3" implementation from https://www.santofortunato.net/resources.

Fig. 3. Two graphlets with the same smallest value for Equation (III.1).

identifier, one chooses to label the nodes of the corresponding subgraph so that the identifier provided by (III.1) is as small as possible.

This method of identifying graphlets is used, amongst many others, in [50], [51] and in the 3-and 4-node graphlet detection routine from [35].

In our study, we focus on 3-node and 4-node graphlets. Since (III.1) can give the same value to different graphlets if they have a different number of nodes, we specify the number of nodes as a prefix of graphlet identifier. That is, the 3-node graphlet identified by 14 will be labelled 3-14, and the 4-node graphlet identified by 14 will be labelled 4-14, as illustrated in Figure 3.

IV. Feature Selection Procedure

In Section V of the main paper, we build a procedure for feature selection on drawing parallels with some basic principles of PCA. We develop below in details how we derive our procedure.

Assuming we have $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_q\end{bmatrix} \in \mathbb{R}^{M\times q}$ the q principal axes of a dataset $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}^T \in \mathbb{R}^{n\times M}$. Denoting by $\mathcal{A}$ the affine space that is the translation of the vector space with basis $\mathbf{U}$ that passes through the gravity centre $\overline{\mathbf{x}} = \frac{1}{n}\left(\mathbf{e}^T\mathbf{X}\right)^T \in \mathbb{R}^M$, with $\mathbf{e}$ a vector of ones. Thus, it is well known that:

- The matrix $\mathbf{U}$ can be thought of in terms of a maximisation of dispersion problem. Indeed, given $q < M$, $\mathbf{U}$ solves

$$\mathcal{P}_q : \begin{cases} \underset{\mathbf{V}\in\mathbb{R}^{M\times q}}{argmax} & \mathcal{I}_{\mathbf{V}}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}||\mathbf{y}_i - \overline{\mathbf{y}}||_2^2 \\ \text{with :} & \mathbf{y}_i = \mathbf{V}\mathbf{V}^T(\mathbf{x}_i - \overline{\mathbf{x}}) + \overline{\mathbf{x}} \\ \text{and :} & \overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \\ \text{subject to :} & \mathbf{V}^T \times \mathbf{V} = \mathbf{I}_q \end{cases} \quad \text{(IV.1)}$$

  That is, the projections of the samples (i.e. the rows) from $\mathbf{X}$ onto $\mathcal{A}$ provide the maximum of inertia[4] $\mathcal{I}_U(\mathbf{X})$ that it is possible to reach by projecting the samples onto an affine space of dimension q passing through $\overline{\mathbf{x}}$.

[4]Given a dataset, the inertia corresponds to the mean square distance between the samples and the mean.

- The projections of samples from $\mathbf{X}$ onto $\mathcal{A}$ are also those that minimise the approximation error, that is $\mathbf{U}$ is also solution of:

$$\mathcal{P'}_q : \begin{cases} \underset{\mathbf{V}\in\mathbb{R}^{M\times q}}{argmin} & \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{y}_i||_2^2 \\ \text{with} & \mathbf{y}_i = \mathbf{V}\mathbf{V}^T(\mathbf{x}_i - \overline{\mathbf{x}}) + \overline{\mathbf{x}} \\ \text{subject to:} & \mathbf{V}^T \times \mathbf{V} = \mathbf{I}_q \end{cases} \quad \text{(IV.2)}$$

- The solution of $\mathcal{P}_{q+1}$ and $\mathcal{P'}_{q+1}$ is of the form $\widehat{\mathbf{U}} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_q & \mathbf{u}_{q+1}\end{bmatrix} = \begin{bmatrix}\mathbf{U} & | & \mathbf{u}_{q+1}\end{bmatrix} \in \mathbb{R}^{M\times(q+1)}$, with $\mathbf{u}_{q+1}$ the $q+1$th principal axis.
- Each principal axis—that is, each column in $\mathbf{U}$—is an eigenvector of the covariance matrix of $\mathbf{X}$: $\mathbf{S}_{\mathbf{X}} = \frac{1}{n}\mathbf{X}_c^T\mathbf{X}_c$. If we denote $\lambda_1, ...\lambda_q \geq 0$ the corresponding eigenvalues, then the maximum of inertia in Eq. (IV.1) is equal to $\mathcal{I}_{\mathbf{U}}(\mathbf{X}) = \sum_{t=1}^{q}\lambda_t$, with each $\mathbf{u}_t, t \in \{1..q\}$ that increases the inertia by λ_t. It is hence possible to sort the principal axes by order of significance by sorting the corresponding eigenvalues.

In the following, our aim is to build a feature selection procedure—that is, picking k axes from the canonical basis—based on the PCA, and that respects similar principles. We assume that we know the q first principal axes of our dataset $\mathbf{X}$, that we denote by $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_q\end{bmatrix}$, and we denote by $\mathbf{Y} \in \mathbb{R}^{n\times M}$ the projection of the rows of $\mathbf{X}$ onto the affine space $\mathcal{A}$ defined above, i.e. $\mathbf{Y} = (\mathbf{X} - \mathbf{e}\times\overline{\mathbf{x}}^T)\mathbf{U}\mathbf{U}^T + \mathbf{e}\times\overline{\mathbf{x}}^T$.

With the PCA principles recalled above as an inspiration, we propose to build our feature selection procedure by resolving

$$\mathcal{Q}_k : \begin{cases} \underset{\mathbf{F}\in\mathbb{R}^{M\times k}}{argmax} & \mathcal{I}_{\mathbf{F}}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}||\mathbf{z}_i - \overline{\mathbf{z}}||_2^2 \\ \text{with :} & \mathbf{z}_i = \mathbf{F}\mathbf{F}^T(\mathbf{y}_i - \overline{\mathbf{y}}) + \overline{\mathbf{y}} \\ \text{and} & \overline{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i \\ \text{subject to :} & \mathbf{F} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots \mathbf{e}_{\sigma_k}\end{bmatrix} \end{cases} \quad \text{(IV.3)}$$

where $\mathbf{e}_j$ is the jth column of the identity matrix. That is, our aim is to maximise the inertia of the projections of samples $\mathbf{y}_i$ onto an affine space, passing through the centre of gravity of $\mathbf{Y}$, and whose vector directions are a subset of the canonical axes.

In the following, we will draw some parallels with PCA principles, and we will derive a score allowing to choose the most promising canonical axes.

We remark that, similarly to PCA principles, the solution of $\mathcal{Q}_k$ is also solution of

$$\mathcal{Q}'_k \begin{cases} \underset{\mathbf{F}\in\mathbb{R}^{M\times k}}{argmin} & \frac{1}{n}\sum_{i=1}^{n}||\mathbf{y}_i - \mathbf{z}_i||_2^2 \\ \text{with} & \mathbf{z}_i = \mathbf{F}\mathbf{F}^T(\mathbf{y}_i - \overline{\mathbf{y}}) + \overline{\mathbf{y}} \\ \text{subject to :} & \mathbf{F} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots \mathbf{e}_{\sigma_k}\end{bmatrix} \end{cases}, \quad \text{(IV.4)}$$

that is, the k canonical axes that maximise the inertia are also those that minimise the approximation error.

We also remark that, as for PCA, the solution of $\mathcal{Q}_{k+1}$ and $\mathcal{Q}'_{k+1}$ can be written as $\widehat{\mathbf{E}} = \begin{bmatrix}\mathbf{E} & | & \mathbf{e}_{\sigma_{k+1}}\end{bmatrix}$.

These two statements are proved below.

Proof. First of all, it is straightforward to observe that $\overline{\mathbf{z}} = \overline{\mathbf{y}} = \overline{\mathbf{x}}$.

Secondly, we show that $\mathcal{Q}_k \iff \mathcal{Q}'_k$:
Let $\mathbf{E} \in \mathbb{R}^{M\times k}$ such that $\mathbf{E} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots \mathbf{e}_{\sigma_k}\end{bmatrix}$:

$$\forall t \in \{1..k\}, \mathbf{e}_{\sigma_t}(i) = \begin{cases} 1 \text{ if } i = \sigma_t \\ 0 \text{ otherwise} \end{cases}. \tag{IV.5}$$

Then

$$\begin{aligned}\mathcal{I}_{\mathbf{E}}(\mathbf{Y}) &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{z}_i - \overline{\mathbf{z}}||_2^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}})\right)^T \left(\mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}})\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{E}\mathbf{E}^T (\mathbf{y}_i - \overline{\mathbf{y}})\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{z}_i||_2^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}}) + \overline{\mathbf{y}} - \mathbf{y}_i\right)^T \\ &\quad \times \left(\mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}}) + \overline{\mathbf{y}} - \mathbf{y}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \big[(\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}}) \\ &\quad -2(\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}}) + ||\mathbf{y}_i - \overline{\mathbf{y}}||_2^2\big] \\ &= -\mathcal{I}_{\mathbf{E}}(\mathbf{Y}) + \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - \overline{\mathbf{y}}||_2^2,\end{aligned}$$

and since $\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - \overline{\mathbf{y}}||_2^2$ does not depend on $\mathbf{E}$, this shows the equivalence.

We now show that the solution of $\mathcal{Q}_{k+1}$ can be written as $\mathbf{E} = \left[\widetilde{\mathbf{E}} \quad | \quad \mathbf{e}_{\sigma_{k+1}}\right]$, with $\widetilde{\mathbf{E}} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots & \mathbf{e}_{\sigma_k}\end{bmatrix}$ solution of $\mathcal{Q}_k$.

We remark that, for a matrix $\mathbf{E}$ subject to Eq. (IV.5), we can rewrite $\mathcal{I}_{\mathbf{E}}(\mathbf{Y})$ as the sum of some diagonal elements of the covariance matrix of $\mathbf{Y}$:

$$\begin{aligned}\mathcal{I}_{\mathbf{E}}(\mathbf{Y}) &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{E}\mathbf{E}^T(\mathbf{y}_i - \overline{\mathbf{y}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{k} \mathbf{e}_{\sigma_t}^T(\mathbf{y}_i - \overline{\mathbf{y}}) \times (\mathbf{y}_i - \overline{\mathbf{y}})^T \mathbf{e}_{\sigma_t} \\ &= \sum_{t=1}^{k} \mathbf{e}_{\sigma_t}^T \left(\frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T\right) \mathbf{e}_{\sigma_t} \\ &= \sum_{t=1}^{k} \mathbf{e}_{\sigma_t}^T \mathbf{S}_{\mathbf{Y}} \mathbf{e}_{\sigma_t} = \sum_{t=1}^{k} \mathbf{S}_{\mathbf{Y}}(\sigma_t, \sigma_t)\end{aligned}$$

where

$$\begin{aligned}\mathbf{S}_{\mathbf{Y}} &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T \\ &= \frac{1}{n}(\mathbf{Y} - \mathbf{e} \times \overline{\mathbf{y}}^T)^T (\mathbf{Y} - \mathbf{e} \times \overline{\mathbf{y}}^T)\end{aligned}$$

is the covariance matrix of the samples in $\mathbf{Y}$, which implies that its diagonal elements are non negative.

Let us assume that we have $\mathbf{E} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots & \mathbf{e}_{\sigma_{k+1}}\end{bmatrix}$ solution of $\mathcal{Q}_{k+1}$. We can write $\mathcal{I}_{\mathbf{E}}(\mathbf{Y}) = \sum_{t=1}^{k+1} \mathbf{S}_{\mathbf{Y}}(\sigma_t, \sigma_t)$. Without loss of generality, we can assume that

$$\mathbf{S}_{\mathbf{Y}}(\sigma_1, \sigma_1) \geq \dots \geq \mathbf{S}_{\mathbf{Y}}(\sigma_k, \sigma_k) \geq \mathbf{S}_{\mathbf{Y}}(\sigma_{k+1}, \sigma_{k+1}) \geq 0$$

Thus, if $\widetilde{\mathbf{E}} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots & \mathbf{e}_{\sigma_k}\end{bmatrix}$ is not a solution of $\mathcal{Q}_k$, it means that it exists a matrix $\mathbf{F} = \begin{bmatrix}\mathbf{e}_{\pi_1} & \dots & \mathbf{e}_{\pi_k}\end{bmatrix}$ subject to Eq. (IV.5) such that

$$\mathcal{I}_{\mathbf{F}}(\mathbf{Y}) = \sum_{t=1}^{k} \mathbf{S}_{\mathbf{Y}}(\pi_t, \pi_t) > \mathcal{I}_{\widetilde{\mathbf{E}}}(\mathbf{Y}) = \sum_{t=1}^{k} \mathbf{S}_{\mathbf{Y}}(\sigma_t, \sigma_t)$$

Hence, denoting by $\widehat{\mathbf{F}}$ the matrix $\widehat{\mathbf{F}} = \begin{bmatrix}\mathbf{F} & | & \mathbf{e}_{\sigma_{k+1}}\end{bmatrix}$, we have $\mathcal{I}_{\widehat{\mathbf{F}}}(\mathbf{Y}) > \mathcal{I}_{\mathbf{E}}(\mathbf{Y})$, which is in contradiction with our assumption that $\mathbf{E}$ is solution of $\mathcal{Q}_{k+1}$.

This proves that if $\mathbf{E} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots & \mathbf{e}_{\sigma_{k+1}}\end{bmatrix}$ such that

$$\mathbf{S}_{\mathbf{Y}}(\sigma_1, \sigma_1) \geq \dots \geq \mathbf{S}_{\mathbf{Y}}(\sigma_k, \sigma_k) \geq \mathbf{S}_{\mathbf{Y}}(\sigma_{k+1}, \sigma_{k+1}) \geq 0$$

is solution of $\mathcal{Q}_{k+1}$, then $\widetilde{\mathbf{E}} = \begin{bmatrix}\mathbf{e}_{\sigma_1} & \dots & \mathbf{e}_{\sigma_k}\end{bmatrix}$ is solution of $\mathcal{Q}_k$. □

With the end of the proof, we observe that, similarly to what happens in PCA, it is possible to write the global inertia $\mathcal{I}_{\mathbf{E}}(\mathbf{Y})$ as the sum of the inertia brought independently by each canonical axis: the pth canonical axis brings $\mathbf{S}_{\mathbf{Y}}(p,p)$ to the inertia. This can be re-written in terms of the largest eigencouples of the covariance matrix $\mathbf{S}_{\mathbf{X}}$ of the initial dataset $\mathbf{X}$. Indeed, recalling that

$$\forall i \in \{1..n\}, \mathbf{y}_i = \mathbf{U}\mathbf{U}^T(\mathbf{x}_i - \overline{\mathbf{x}}) + \overline{\mathbf{x}},$$

with $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_q\end{bmatrix}$, such that $\forall t, \mathbf{u}_t$ is the normalised eigenvector of $\mathbf{S}_{\mathbf{X}}$ associated with λ_t, and $\lambda_1 \geq ... \geq \lambda_q \geq 0$.

Let $k \in \{1...M\}$, then:

$$\begin{aligned}
\mathbf{S_Y}(p,p) &= \frac{1}{n}\mathbf{e}_p^T \left(\sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T \right) \mathbf{e}_p \\
&= \frac{1}{n}\mathbf{e}_p^T \left(\sum_{i=1}^{n} (\mathbf{U}\mathbf{U}^T(\mathbf{x}_i - \overline{\mathbf{x}}))(\mathbf{U}\mathbf{U}^T(\mathbf{x}_i - \overline{\mathbf{x}}))^T \right) \mathbf{e}_p \\
&= \mathbf{e}_p^T \mathbf{U}\mathbf{U}^T \left(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \right) \mathbf{U}\mathbf{U}^T \mathbf{e}_p \\
&= \mathbf{e}_p^T \mathbf{U}\mathbf{U}^T \mathbf{S_X} \mathbf{U}\mathbf{U}^T \mathbf{e}_p \\
&= \mathbf{e}_p^T \mathbf{U} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_q \end{bmatrix} \mathbf{U}^T \mathbf{e}_p \\
&= \begin{bmatrix} \mathbf{u}_1(p) & \dots & \mathbf{u}_q(p) \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_q \end{bmatrix} \begin{bmatrix} \mathbf{u}_1(p) \\ \vdots \\ \mathbf{u}_q(p) \end{bmatrix} \\
&= \sum_{t=1}^{q} \lambda_t \mathbf{u}_t(p)^2,
\end{aligned}$$

with $\mathbf{e}_p$ the pth column of the identity matrix.

Hence, we can derive a score that assesses the significance of a canonical axis, given the q first principal axes:

Definition 1. *The expression level of the kth canonical axis within the q first principal axes is defined to be*

$$\gamma(p) = \frac{1}{\sum_{s=1}^{q} \lambda_s} \sum_{t=1}^{q} \lambda_t \mathbf{u}_t(p)^2, \tag{IV.6}$$

The measure $\gamma(p)$ gives an indication of the significance of the pth canonical axis, and is normalised so that $\sum_{p=1}^{k} \gamma(p) = 1$.

Remark 1. *If the number of principal axes taken into account is equal to the size of samples, then $\gamma(p) = \mathbf{S_X}(p,p), \forall p$.*

V. Setting and Parameters used for Numerical Experiments

In this section, we justify and develop our choice of parameters for the numerical experiments, corresponding to Sections IV-A and V-A from the main paper.

A. Number of Principal Components kept in the Feature Extraction Procedure

In Section IV-A of the main paper, we choose to keep 11 principal components. Our rationale for choosing this number comes from the following observations. We have measured the $F1$-score of our method for several number of kept principal axes (2 to 19). The error bars for each benchmark is presented in Figure 4. Vertical bars indicate a range of plus or minus one standard deviation around the average $F1$-score, given a fixed number of principal components kept in the feature extraction procedure. In this figure, we observe that a globally stable situation is reached when at least 5 principal components are kept, for each benchmark.

If one zooms on $F1$-scores for 5-or-more kept components, as in Figure 5[5], one can observe that, for food webs and social networks, there is the gap between the $F1$-score if one keeps 11 components and $F1$-score if one keeps less than 11 components (this threshold being quite sharp for social networks). This is not the case for electronic circuits, where keeping exactly 11 components even leads to an average $F1$-score that one can qualify as a local minimum. On the other hand, the standard deviation for this benchmark is sensibly higher than for the others, the shape of its $F1$-score curve is less clear. For rhetorical discourse structures, whereas keeping 11 components does not lead to a local maximum or a threshold, it is nevertheless at the middle value of some increasing slope.

B. Parameters for `graph2vec` *and* `gl2vec`

Given a network G, the embedding method `gl2vec` compares the number of occurrences of 3-node graphlets in G with the expected number of 3-node graphlets in random graphs sharing some characteristics with the initial networks [52]. The authors of [52] compare three random models and conclude that the best one is the one that generates random graphs with number of edges and nodes equal to those of G, with no other constraint. This is thus the random model we have chosen in our experiments. No other parameters needs to be set up for this method.

On the other hand, `graph2vec` [53] required more investigation. First, as it works for undirected networks only we provide an undirected version of networks by not taking into account the direction of the edges. As node labels (required in Weisfeiler-Lehman kernels on which `graph2vec` is based), we provided a pair containing the in- and out-degrees of each node in the directed network. The dimension of embedding and the maximum degree of subgraphs to consider in the kernels are parameters to set up. We did a grid search for embedding size in $\{2^k, k = 1..9\}$ and kernel depth from 1 to 4. We kept the most accurate results, which falls for embedding of size 64 and the kernel depth equal to 1.

C. Parameters for Feature Selection Based on Random Forests

As a matter of comparison for our feature selection procedure presented in Section V from the main paper, we have confronted it against another feature selection procedure based on random forests (RF) [54]. We explain below how we processed.

First, we have embedded the networks $G_i, \forall i \in \mathcal{I}$ to obtain their vector representations $\mathbf{x}_i, \forall i \in \mathcal{I}$:

$$\forall i \in \mathcal{I}, \quad \mathbf{x}_i = \frac{\bigoplus_{k=3,4} \alpha_k(G_i)\phi_k(G_i)}{\sqrt{\sum_{k=3,4} \alpha_k(G_i)^2 \phi_k(G_i)^T \phi_k(G_i)}}, \tag{V.1}$$

where $\phi_k(G) \in \mathbb{R}^{m_k}$ is the vector such that $\phi_k(G)(p)$ is the number of occurrences of the pth k-node graphlet in G,

[5] In this figure, the error bars have been divided by 10 to improve the visualisation.

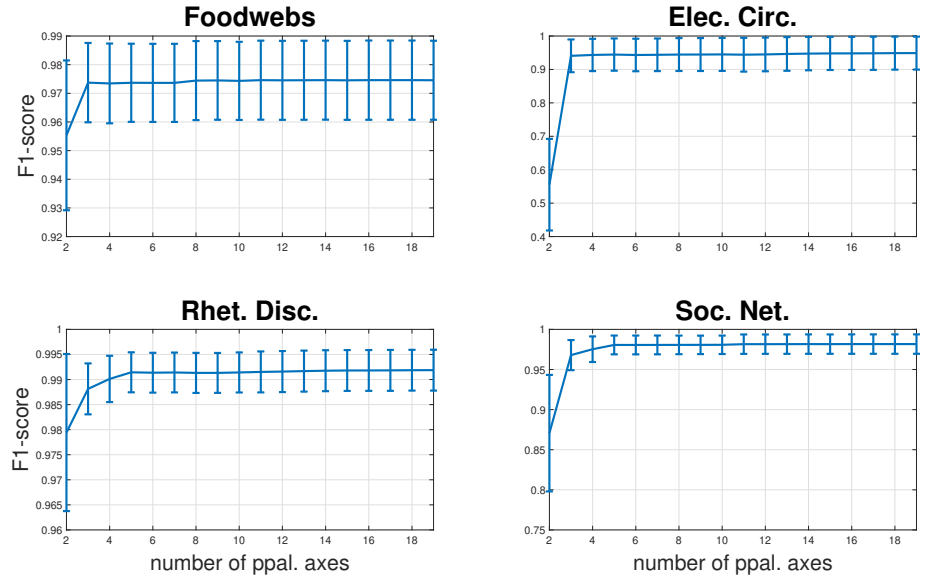

Fig. 4. F1-score on each benchmark, according to the number of components that are kept

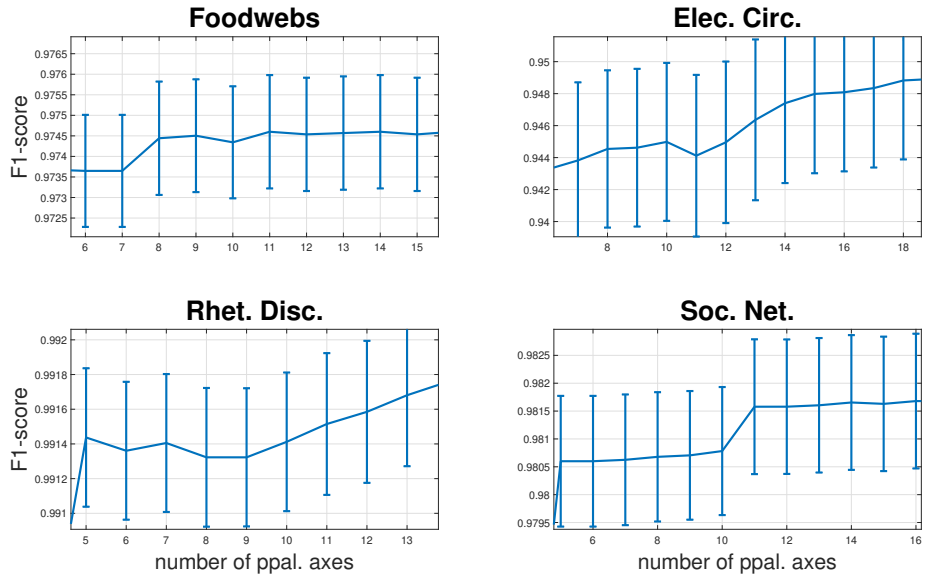

Fig. 5. Zoom on F1-score on each benchmark, according to the number of components that are kept

and the factors $\alpha_k(G) = \frac{m_k}{k}\binom{k}{n}^{-1}$, with n the number of nodes in G and m_k the number of non-isomorphic k-node connected graphlets[6], are used to mitigate the order of magnitude difference between $\phi_3(G)$ and $\phi_4(G)$. The symbol $\bigoplus$ provides the direct sum of the involved vectors.

We have chosen to apply the RF-based feature selection on network embeddings provided by Eq. (V.1) because these are also the embeddings we use in our own feature selection procedure. On the other hand, we have tried to apply RF-based feature selection simply on $\bigoplus_{k=3,4} \phi_k(G)$ (normalised or not), and the results for the RF-based feature selection were much worse than those we present in the main paper.

To measure a RF-based feature significance,for each training set $\mathcal{S}$, we have trained a RF and recovered the Gini importance—or mean decrease in impurity—of each graphlet according to the forest [55]. The Gini importance of a forest being the average Gini importance over all trees. We have used the random forest implementation proposed in the python

[6] $m_k = 13$ if $k = 3$ and $m_k = 212$ if $k = 4$

library `scikit-learn` [56] with all default parameters—essentially, 100 trees are built in the forest, and the Gini impurity is used to measure the quality of a split in a tree.

We have kept the 9 graphlets with highest Gini importance, that is, the same number of graphlets we kept for our procedure.

VI. Remark

Tables and Figures from this paper and [1] can be reproduced by using the Matlab code and dataset at http://github.com/luleg/DiscriminantMotifs, where a curated version of the networks detailed in Section II can also be found.

References

[1] L. le Gorrec and P. A. Knight, "A simple embedding for classifying networks with a few graphlets," 2020. Submitted to ASONAM 2020.

[2] https://www.globalwebdb.com/. Access: February 2020.

[3] J. Dunne, K. Lafferty, A. Dobson, R. Hechinger, A. Kuris, N. Martinez, J. McLaughlin, K. Mouritsen, R. Poulin, K. Reise, D. Stouffer, D. Thieltges, R. Williams, and C. Zander, "Parasites affect food web structure primarily through increased diversity and complexity," *PLoS biology*, vol. 11, no. 6, 2013.

[4] C. D. Zander, N. Josten, K. C. Detloff, R. Poulin, J. P. McLaughlin, and D. W. Thieltges, "Food web including metazoan parasites for a brackish shallow water ecosystem in germany and denmark: Ecological archives e092-174," *Ecology*, vol. 92, no. 10, pp. 2007–2007, 2011.

[5] V. Batagelj and A. Mrvar, "Pajek datasets." http://vlado.fmf.uni-lj.si/pub/networks/data/, 2006. Access: March 2020.

[6] J. Kunegis, "KONECT – The Koblenz Network Collection," in *Proc. Int. Conf. on World Wide Web Companion*, pp. 1343–1350, 2013.

[7] N. D. Martinez, "Artifacts or attributes? effects of resolution on the little rock lake food web," *Ecological monographs*, vol. 61, no. 4, pp. 367–392, 1991.

[8] A. Clauset, E. Tucker, and M. Sainz, "The Colorado index of complex networks." https://icon.colorado.edu/, 2016. Access: February 2020.

[9] F. Brglez, D. Bryan, and K. Kozminski, "Combinational profiles of sequential benchmark circuits," in *IEEE International Symposium on Circuits and Systems*, pp. 1929–1934, 1989.

[10] F. Corno, M. S. Reorda, and G. Squillero, "Rt-level itc'99 benchmarks and first atpg results," *IEEE Design & Test of computers*, vol. 17, no. 3, pp. 44–53, 2000.

[11] http://pld.ttu.ee/~maksim/benchmarks/iscas89/bench/. Access: February 2020.

[12] N. Asher and A. Lascarides, *Logics of Conversation*. Cambridge University Press, 2003.

[13] A. Lascarides and N. Asher, *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*, pp. 87–124. Springer, 2007.

[14] M. Qi, "Use of deep learning approaches for the prediction of discourse structures," Master's thesis, Université Paul Sabatier, Toulouse, 2018.

[15] N. Asher, J. Hunter, M. Morey, F. Benamara, and S. Afantenos, "Discourse structure and dialogue acts in multiparty dialogue: the stac corpus," in *LREC*, pp. 2721–2727, 2016.

[16] J. Moody, "Peer influence groups: Identifying dense clusters in large networks," *Social Networks*, vol. 23, pp. 261–283, 10 2001.

[17] http://moreno.ss.uci.edu/data.html. Access: February 2020.

[18] J. S. Coleman, *Introduction to Mathematical Sociology*, pp. 450–451. New York: Free Press, 1964.

[19] C. M. W. L. C. Freeman and D. M. Kirke, "Exploring social structure using dynamic three-dimensional color images," *Social Networks*, vol. 20, pp. 109–118, 1998.

[20] C. Baerveldt, "Pupil's networks in high schools. network sampling, program and some results from a theory-oriented research project on petty crime of pupils," in *Paper for the 2nd International Network Sampling Workshop*, 2000.

[21] http://www.stats.ox.ac.uk/~snijders/siena/BaerveldtData.html. Access: March 2020.

[22] E. Lazega, *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press, 2001.

[23] J. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, vol. 20, no. 4, pp. 253–270, 1957.

[24] R. Cross and A. Parker, *The Hidden Power of Social Networks*. Harvard Business School Press, Boston, MA, 2004.

[25] http://toreopsahl.com/datasets. Access: March 2020.

[26] M. R. Weeks, S. Clair, S. P. Borgatti, K. Radda, and J. J. Schensul, "Social networks of drug users in high-risk sites: Finding the connections," *AIDS and Behaviour*, vol. 6, pp. 193–206, 2002.

[27] http://sites.google.com/site/sfeverton18/research/cohesion-and-clustering. Access: March 2020.

[28] M. J., "Direct factor analysis of sociometric data," *Sociometry*, vol. 23, pp. 360–371, 1960.

[29] C. P. Loomis, J. O. Morales, R. A. Clifford, and O. E. L. Turrialba, *Social Systems and the Introduction of Change*. Glencoe (Ill.): The Free Press, 1953.

[30] . V. B. W. de Nooy, A. Mrvar, *Exploratory Social Network Analysis with Pajek*, ch. 3. Cambridge: Cambridge University Press, 2004.

[31] M. Fire, R. Puzis, and Y. Elovici, "Link prediction in highly fractional data sets," *Handbook of Computational Approaches to Counterterrorism*, 2012.

[32] http://proj.ise.bgu.ac.il/sns/. Access: March 2020.

[33] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The slashdot zoo: Mining a social network with negative edges," in *Proceedings of the 18th International Conference on World Wide Web*, p. 741–750, 2009.

[34] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[35] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." http://snap.stanford.edu/data, June 2014. Access: February 2020.

[36] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 933–943, International World Wide Web Conferences Steering Committee, 2018.

[37] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," in *The Semantic Web-ISWC*, pp. 351–368, Springer, 2003.

[38] P. Massa, M. Salvetti, and D. Tomasoni, "Bowling alone and trust decline in social network sites," in *Proc. Int. Conf. Dependable, Autonomic and Secure Computing*, pp. 658–663, 2009.

[39] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos, "Edge weight prediction in weighted signed networks," in *Data Mining (ICDM), 2016 IEEE International Conference on*, 2016.

[40] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, (New York, NY, USA), p. 333–341, Association for Computing Machinery, 2018.

[41] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *ACM Workshop on Online Social Networks*, (New York, NY, USA), p. 37–42, Association for Computing Machinery, 2009.

[42] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 555–564, 2017.

[43] R. Michalski, S. Palus, and P. Kazienko, "Matching organizational structure and social network extracted from email communication," in *Int. Conf. on Bus. Inf. Syst.*, pp. 197—-206, 2011.

[44] D. Homscheid, J. Kunegis, and M. Schaarschmidt, "Private-collective innovation and open source software: Longitudinal insights from linux kernel development," in *Open and Big Data Management and Innovation* (M. Janssen, M. Mäntymäki, J. Hidders, B. Klievink, W. Lamersdorf, B. van Loenen, and A. Zuiderwijk, eds.), (Cham), pp. 299–313, Springer International Publishing, 2015.

[45] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Soc. Netw.*, vol. 31, pp. 155—-163, 2009.

[46] P. Sapiezynski, A. Stopczynski, D. Lassen, and S. Lehmann, "Interaction data from the Copenhagen Networks Study," *Scientific Data*, vol. 6, no. 1, 2019.

[47] C. Morselli, *Inside criminal networks*. New York: Springer, 2009.

[48] http://sites.google.com/site/ucinetsoftware/datasets/covert-networks/caviar. Access: March 2020.

[49] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Phys. Rev. E*, vol. 80, p. 016118, Jul 2009.

[50] A. Topirceanu, A. Duma, and M. Udrescu, "Uncovering the fingerprint of online social networks using a network motif based approach," *Computer Communications*, vol. 73, pp. 167–175, 2016.

[51] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.

[52] K. Tu, J. Li, D. Towsley, D. Braines, and L. D. Turner, "Gl2vec: Learning feature representation using graphlets for directed networks," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, (New York, NY, USA), p. 216–221, Association for Computing Machinery, 2019.

[53] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," *CoRR*, vol. abs/1707.05005, 2017.

[54] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?," *Bioinformatics*, vol. 34, pp. 3711–3718, 05 2018.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.