Network analysis improves interpretation of affective physiological data

Yuriy Hulovatyy[1], Sidney D'Mello[1,2], Rafael A. Calvo[3], and Tijana Milenković[1,*]

[1]Dept. of Computer Science and Engineering, University of Notre Dame; [2]Dept. of Psychology, University of Notre Dame
[3]Dept. of Electrical Engineering, University of Sydney; [*]Corresponding Author (E-mail: tmilenko@nd.edu)

***Abstract*—Understanding how human physiological responses to a given stimulus vary across individuals is critical for the fields of Affective Psychophysiology and Affective Computing. We approach this problem via network analysis. By analyzing individuals' galvanic skin responses (GSRs) to a set of emotionally charged images, we model each image as a network, in which nodes are individuals and two individuals are linked if their GSRs to the given image are similar. In this context, we evaluate several network inference strategies. Then, we group (or cluster) images with similar network topologies, while evaluating a number of clustering choices. We compare the resulting network-based partitions against the known arousal/valence-based "ground truth" partition of the image set (which is likely noisy). While our network-based image partitions are statistically significantly similar to the "ground truth" partition (meaning that network analysis correctly captures the underlying signal in the data), the network-based partitions outperform the "ground truth" partition with respect to an independent criterion, namely in terms of latent semantic analysis (meaning that our partitions are more semantically meaningful than the "ground truth" partition). Thus, network analysis of affective physiological data appears to improve interpretation of the data. To our knowledge, we are the first to use a network-based approach to study this data type.**

I. Introduction

Networks (or graphs) model interactions between elements of a system. They enable the study of complex processes that emerge from the collective behavior of interconnected elements. Hence, networks have been useful models for real phenomena in numerous domains, e.g., social, technological, or biological systems [1], [2], [3]. We focus on network modeling of affective physiological data, in order to gain insights into how individuals physiologically respond to emotional stimuli, thereby benefiting the fields of Affective Psychophysiology and Affective Computing. To our knowledge, we are the first to use network analysis to study affective physiological data, which, as we will show, improves the interpretation of the data compared to an alternative non-network-based approach.

A. Motivation and background

There is an inextricable coupling between physiology and emotions because one of the key evolutionary functions of emotion is to facilitate rapid action in response to relevant environmental events [4]. Emotions are constructs (or conceptual entities) that cannot be directly measured, but must be inferred from measurable signals like physiology. Therefore, understanding the relationship between emotions and physiology has been an important endeavor in the field of Affective Psychophysiology for more than a century.

There is also an engineering side to complement the scientific endeavor of identifying the physiological correlates of affect. The field of Affective Computing, which is a subfield of Human-Computer Interaction, aims to build intelligent systems that respond to user emotions much like an actual human would [5]. For example, a system can offer a hint if it detects that a user is confused or frustrated. Considerable work has focused on developing automated approaches to detect emotions from observable signals like facial expressions, speech patterns, etc. (see [6], [7] for recent reviews). Physiological-based approaches for affect detection are attractive (e.g., [8], [9], [10]), because these signals are largely involuntary and thereby less susceptible to social masking like facial expressions and speech. This once again raises the fundamental issue of understanding the relationship between affect and physiology, which is the focus of this paper.

B. Related work

Over the last century, many attempts have been made to identify how different emotions are manifested in physiological signals, such as the electrocardiogram (ECG), electromyogram (EMG), or galvanic skin response (GSR) (see [11] for a review). While it was once claimed that unique discrete emotions (e.g., fear, anger) are accompanied by distinct physiological patterns [12], meta-analyses and other syntheses of the literature have failed to conclusively support this claim [11].

One reason for the difficulties in uncovering the emotion → physiological mapping is the considerable individual variability in emotional responses (i.e., reactions to the same stimulus vary across individuals). In affect detection, the most common approach to handle this variability has been to simply ignore it (e.g., [9], [13], [14]). This is done by building *person-dependent* models that are carefully calibrated to each individual. However, despite some advances [8], [15], what is needed are *person-independent* models that generalize to new individuals. The few efforts along this front have produced mixed results [16]. For example, recognition-rates of person-dependent and person-independent models from three physiological signals (ECG, EMG, and GSR) were compared [10], and machine learning applied to detect seven emotions indicated that it was possible to build person-dependent models with moderate classification accuracy, but accuracy went to zero when person-independent modeling was adopted.

C. Our approach

To summarize, efforts to identify unique emotion-specific physiological responses have been largely unsuccessful, likely

due to considerable intra- and inter- individual variability in the signals. Unfortunately, most (but not all) of the research have typically considered this variability to be sources of error and something to be averaged over. In our view, however, this variability is far from random as there might be structure in the noisy patterns of individual responding. Modeling this variability will provide insights into the fundamental question of how individuals physiologically respond to emotions.

This paper adopts a novel approach to study interrelations among individuals by formulating the problem from a network perspective. Since networks model relationships between objects, and since physiological variability depends on the interrelations – both between humans and their physiological states – by looking at the system of stimuli and individual responses to them from a structural (or topological) point of view, networks can provide important insights on the problem of modeling and understanding this variability.

To this end, we consider affective physiological responses of humans to a set of stimuli and construct networks reflecting relations between individual responses. We then analyze and compare these networks to investigate how inter-individual patterns of responses map onto the known "ground truth" about the stimuli. In addition to this, we provide a comprehensive analysis of the ways to construct, compare, and cluster networks, which allows us to examine relative effects of choosing different methods and parameters. As a result, we develop an extensible framework for systematic analysis of affective physiological data. Also, we demonstrate that network analysis of such data improves the interpretation of the data compared to an alternative, non-network-based approach.

II. Methods

The experimental setup for collection of affective physiological data used in this study is described in Section II-A. From this data, we construct networks using a variety of network inference strategies (Section II-B). We evaluate topological properties of the networks (Section II-C), in order to group (or cluster) networks that have similar properties. For this purpose, we use several clustering methods and various combinations of their parameters (Section II-D). We conduct a thorough evaluation of the quality of partitions produced by the different clustering strategies and verify both statistical and practical significance of our results (Section II-D4).

A. Data collection

Eighteen human subjects were presented with 89 emotionally charged images from the International Affective Picture System (IAPS) [17]. The IAPS is a collection of over a thousand images depicting people, objects, or events that have been selected on the basis of how they evoke valence (unpleasant to pleasant) and arousal (sleepy to active) in large samples of viewers. Arousal and valence are the two fundamental dimensions of affective responses [18].

Each subject was equipped with a sensor to record their galvanic skin (or electrodermal) response (GSR). GSR tracks the electrical conductivity of the skin based on variations in moisture caused by sweating. The basic idea of GSR is that the sweat glands are controlled by the sympathetic nervous system (which modulates affect related flight or fight responses), so variations in moisture that are picked up by the GSR signal can reflect changes in physiological arousal.

A GSR signal of each subject to each image was obtained, resulting in a total number of $18 \times 89 = 1,602$ signals. Each subject viewed each image for 10 seconds, with responses being recorded at a rate of 1,000 Hz. The 89 signals for each subject were first standardized (converted to z scores) within the subject and then smoothed with a 0.3 Hz low-pass filter.

The 89 images were selected in a way that the IAPS arousal and valence normative scores [17] for the stimuli spanned a 3×3 arousal/valence space, with the three arousal/valence values corresponding to "low", "medium", or "high". Based on this space, we can partition the images into one of the nine classes (i.e., high arousal - high valence, high arousal - medium valence, high arousal - low valence; medium arousal - high valence, and so on). Henceforth, we refer to this nine-class image partition as the *arousal/valence (AV) "ground truth" partition* (or simply as the *AV partition*). Eight of the nine classes contain 10 images, and one class contains nine images.

Note that we intentionally use quotes when talking about the AV partition, since the "ground truth" is itself quite noisy. This is because the "ground truth" was obtained from a combination of theory (images were carefully selected to evoke particular responses; e.g., spider to evoke fear, or surgery for disgust) and normative ratings of valence and arousal that accompany the IAPS collection. These ratings reflect the *average* valence and arousal as *subjectively* reported by a large sample of individuals (different from our 18 subjects) after viewing each image. Therefore, the assignment of the images to the nine "ground truth" classes in the AV partition could be noisy, which is typical for any emotion-eliciting stimuli.

B. Network construction

We construct networks from the collected physiological response data as follows (Fig. 1). For each image, we create an unweighted, undirected network in which a node corresponds to a subject and an edge exists between two nodes if GSR responses of the corresponding subjects are "similar enough". By "similar enough", we mean that the similarity of two given responses, computed with respect to a given similarity measure, is above a given threshold. Thus, each network structurally captures how different subjects respond to the same image as an initial step of modeling inter-individual variability.

To test whether the choice of the similarity measure and threshold value affects the results, we evaluate two different similarity measures: 1) Pearson correlation (PC) and 2) mutual information (MI) (see below). Then, for each similarity measure, we examine multiple similarity thresholds; we discuss the choice of an appropriate threshold in Section III-A.

When we use PC, we construct networks in two ways. First, we add an edge between two subjects if the PC coefficient between the subjects' responses to the corresponding image is above a given positive threshold. In such a network, only subjects whose responses are strongly positively correlated are linked. We refer to networks constructed in this way as *PC positive networks*. Second, we add an edge between two subjects if the *absolute value* of PC coefficient between the subjects' responses to the given image is above a given

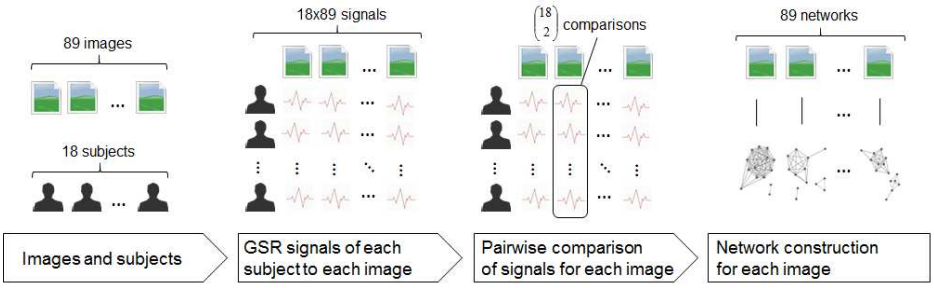


Fig. 1: Summary of the network construction procedure. First, we obtain GSR signals of 18 subjects to 89 images, resulting in 18×89 GSR signals. Then, for each image, we compare responses between each pair of subjects, resulting in $\binom{18}{2}$ comparisons. Finally, for each of the 89 images, we construct a network linking "highly similar" pairs of subjects. Clearly, if the subjects' GSR signals to two (or more) images are similar, then networks corresponding to the images will be "topologically similar".

threshold. In such a network, both subjects whose responses are strongly positively correlated as well as subjects whose responses are strongly negatively correlated are linked. We refer to networks constructed in this way as *PC positive-negative networks*. We do not create networks in which PC coefficient between the subjects' responses to the corresponding image is below a given negative threshold, as such networks would be topologically constrained (in terms of the number of triangles) and thus structurally different from real-world networks [3].

When we use MI, we add an edge between two subjects if the normalized MI [19] between the subjects' responses to the given image is above a given threshold. Compared to PC-based networks, MI-based networks can capture more complex (i.e., non-linear) relationships between responses [19], [20]. Since the normalized MI can have only positive values, we refer to networks constructed in this way as *MI positive networks*.

Thus, we create three different network types corresponding to the two similarity measures: PC positive, PC positive-negative, and MI positive networks. For each network type, for each similarity threshold, we generate 89 networks, one for each of the 89 images. Each of the networks models similarity in responses between the 18 subjects and can thus have 18 nodes. However, since isolated nodes (i.e., nodes with no edges adjacent to them) do not contribute to the topology of the network, we remove such nodes. Therefore, some networks may end up having fewer than 18 nodes (Section III-B).

C. Network properties

To study whether network structure varies across the different network types, we analyze *six* popular concise "footprints" of network structure, or *network properties*: the number of nodes, number of edges, average degree, size of the largest connected component, diameter (see below), and the average clustering coefficient (see below). The diameter of a network is the maximum of shortest path lengths over all pairs of nodes in the network. The average clustering coefficient of a network is the average of clustering coefficients over all nodes in the network; the clustering coefficient of a node is the probability that two of the node's neighbors are connected [21].

D. Clustering

Upon constructing the networks and evaluating their structure, we next ask whether networks corresponding to images of the same AV "ground truth" class (Section II-A) are more "topologically similar" than networks corresponding to images from different classes. To answer this, we cluster the networks into non-overlapping groups based purely on their topological similarities, without using any "ground truth" knowledge about which network (i.e., image) corresponds to which "ground truth" class. In this way, we produce a *network-based* partition of the images. Then, we can compare such a partition with the AV "ground truth" partition (Section II-A), in order to determine whether the two partitions significantly overlap. A significant overlap would indicate that based solely on network topological similarity we can group together images which group together according to the "ground truth". (Note: we do *not* perform "graph clustering" of an *individual network* into groups (or communities) of nodes (or edges) [22], [23].)

To cluster a set of objects (i.e., images), one needs to define: 1) a measure of distance (or equivalently, similarity) between the objects; 2) a clustering method; and 3) parameters of the clustering method. We comprehensively test multiple network-based (as well as non-network-based) distance measures (Sections II-D1 and II-D2), clustering methods (Section II-D3), and variations of the methods' parameters (Section II-D3). Then, upon producing a partition of images, one needs to evaluate the quality of the partition. We evaluate our partitions with respect to their overlap with the "ground truth" knowledge about the images (Section II-D4), as follows.

*1) **Network-based distance measures**:* We use *seven* network similarity measures: 1) *common edges*, i.e., the overlap of the networks' edge sets, as measured by Jaccard index ($\frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$, where E_1 and E_2 are the two edge sets) [24]; 2) absolute *difference of* the networks' *average clustering coefficients* [25]; 3) absolute *difference of* their *average diameters* (the average diameter of the network is the average of shortest path lengths over all node pairs [25]); 4) *Pearson correlation of* the networks' *degree distributions* [25], [3]; 5) *Pearson correlation of* their *clustering spectra* (the clustering spectrum of a network is the distribution of average clustering coefficients of nodes with a particular degree) [25]; 6) relative graphlet frequency distance (*RGF-distance*) (which compares frequencies of all 3-5-node subgraphs, or*graphlets*, in two networks [26]); and 7) graphlet degree distribution agreement (*GDD-agreement*) (which generalizes the degree distribution into a spectrum of graphlet degree distributions [27]). We use *GraphCrunch* for all comparisons [21], [25].

RGF-distance and GDD-agreement take into account more of network topology compared to the other measures, and are thus considered to be more constraining measures of network similarity [26], [27], [28]. Therefore, one would expect that RGF-distance and GDD-agreement would outperform the less constraining measures in terms of partition quality. Surprisingly, as we show in Section III-C, this is not the case.

The common edges measure explicitly takes into account correspondence of node labels between networks, whereas all other measures are insensitive to label correspondence and consider only topological information. So, one would expect that common edges would outperform other measures. Indeed, as we show in Section III-C, this is what we observe in general.

2) ***Non-network-based distance measures****:* Importantly, we want to ensure that we can obtain a more precise image partition by clustering network-based representations of the images via a network similarity measure than by clustering a non-network representation of the image data via some statistical, non-network-based image similarity measure. That is, we want to ensure that it is indeed beneficial to use network analysis to study physiological data. For this purpose, we define an additional non-network-based measure of similarity between two images, as follows. For each image and for each similarity measure (i.e., PC or MI; Section II-B), we construct a vector with $\binom{18}{2}$ values, where each value corresponds to similarity between responses of two corresponding subjects to the given image. Now, instead of selecting a similarity threshold to define a subset of the $\binom{18}{2}$ values as edges in a network corresponding to the given image and then computing similarity between two images by comparing their networks, as above, here, to compute similarity between two images, we directly compare two $\binom{18}{2}$-dimensional vectors corresponding to two images using Pearson correlation. We refer to this measure as the *non-network (NON) distance*.

Thus, we deal with the total of *eight* distance measures. For each, we construct matrices of pairwise image distances and input these matrices into a clustering method (see below), in order to group (separate) similar (dissimilar) images.

3) ***Clustering methods and their parameters****:* To test the effect of the choice of clustering method on partition quality, we use two clustering methods: hierarchical clustering (HIE) and k-medoids clustering (KM) [22]. Parameters of HIE are: the linkage method, i.e., the way of measuring the distance between two clusters, and the choice of how to cut a hierarchical tree (or dendrogram). Clearly, different values of a parameter could lead to different partitions. Thus, we test *four* linkage methods: *single*, *complete*, *average*, and *weighted* [22]. To cut a dendrogram, we specify the desired number of clusters, k, testing all possible values of k, from 1 to 89, in increments of 1. Note that we also tested other strategies for cutting a dendrogram (e.g., by specifying the maximum allowed inter-cluster distance), but their performance in terms of the partition quality was inferior (results not shown). KM is a version of k-means in which cluster centers are data points [29]. The only parameter of KM is the desired number of clusters, k, for which we again test all 89 possible values.

For each network type, given 7 network-based distance measures, 4 linkage methods, and 89 values of k for HIE and 7 network-based distance measures and 89 values of k for KM, we produce $(7 \times 4 \times 89) + (7 \times 89) = 2,492 + 623 = 3,115$ network-based partitions. Since we consider three network types (Section II-B), we produce the total of $3 \times 3,115 = 9,345$ network-based partitions. Also, by using the NON distance, for each measure of similarity between two responses, given 4 linkage methods and 89 values of k for HIE and 89 values of k for KM, we produce $(4 \times 89) + (89) = 356 + 89 = 445$ non-network based partitions. Since we consider two response similarity measures (Section II-B), we produce the total of $2 \times 445 = 890$ such partitions. So, we produce $9,345 + 890 = 10,235$ partitions, each of which we evaluate as follows.

4) ***Quality of partitions****:* To compare any two partitions, we use a measure called *Adjusted for chance Information Distance* (AID) [20], [30]. This measure uses notions of entropy and MI to determine the similarity between two partitions from an information theoretic perspective. Intuitively, it quantifies how much knowing one of the partitions reduces uncertainty about the other [20]. The lower the AID value, the more similar two partitions. AID already incorporates "adjustment for chance" that allows for comparing partitions of different cluster sizes without bias [30]. As a consequence, AID gives a way to rank pairs of partitions based on their similarities. This is very useful in our study, because it allows us to evaluate the fit of the AV "ground truth" partition to many different partitions resulting from the different clustering strategies. And by comparing multiple partitions to the AV partition, we can determine which one of them is closer to the "ground truth".

To determine statistical significance of an AID score, we generate 10,000 random partitions and compute AID between each random partition and the AV partition, resulting in the total of 10,000 "random" AIDs. Then, we determine p-value of our actual AID score as the number of AIDs out of the 10,000 "random" AIDs that have the same or better (i.e., lower) value than the actual AID score. We use p-value threshold of 0.01.

In addition to comparing our network- and non-network-based partitions with the AV "ground truth" partition, we also assess them using latent semantic analysis (LSA) [31]. LSA is a statistical technique that computes the conceptual similarity of two texts (words, sentences, or documents) by leveraging second-order co-occurrence relationships from large text corpora. By considering LSA-derived similarity between descriptions of objects depicted in pairs of IAPS images (e.g., "dog" and "spider"), we obtain pairwise semantic similarity between the images (the online LSA tools (http://lsa.colorado.edu/) were used for the requisite computation). That is, we obtain an additional "ground truth" data set that is based purely on semantic meanings of the images. Now, to evaluate the quality of a partition from the LSA perspective, we compare intra- and inter-cluster LSA similarities within this partition. In a good partition, intra-cluster similarities would be statistically significantly higher than inter-cluster similarities. To evaluate the statistical significance of the difference between intra- and inter-cluster similarities in a partition, we compare the vector of all pairwise intra-cluster similarities with the vector of all pairwise inter-cluster similarities using the Wilcoxon rank-sum test [32]. By doing this for multiple partitions, and by comparing the resulting p-values across the partitions, we can determine which of the partitions is more significant and thus more semantically meaningful. We perform this not only for our network-based partitions but also for the AV "ground truth" partition (which might be noisy; Section II-A), in order to evaluate whether any of our network-based partitions improve upon the AV partition with respect to LSA.

III. Results and Discussion

We describe the choice of parameter values for network construction (Section III-A). We discuss trends in topology

of the constructed networks and the effect of the choice of network construction strategy on the trends (Section III-B). We present image clustering results and discuss the effect of the choice of clustering parameters on the results (Section III-C). We conclude with open research questions (Section IV).

A. Choosing network construction parameters

We use two measures of similarity between GSR signals: Pearson correlation (PC) and mutual information (MI) (Section II-B). Given an image and a response similarity measure, we connect two subjects by an edge if the similarity of their GSR signals is above a given threshold (Section II-B). Ideally, one should select a threshold that is statistically significant. Also, the threshold should provide a meaningful representation as well as interpretation of the data. In this context, we aim to construct a network that ideally links all 18 subjects, in order to include into the network as much of information from the data as possible. At the same time, we aim to construct a network that is not too dense, in order to mimic the sparse nature of many real-world networks as well as avoid randomness in network topology [3]. The choice of threshold value is likely to affect the resulting network topology: the higher the threshold, the smaller the number of nodes included into the network but the smaller the density; the lower the threshold, the larger the number of nodes included into the network but the larger the density. To determine an appropriate threshold, we analyze trends in network topology at multiple thresholds, as follows.

First, we focus on PC positive networks (Section II-B). We vary PC threshold from 0.5 to 0.9 in increments of 0.1 and further from 0.9 to 0.99 in finer increments of 0.01. We do not examine thresholds below 0.5, as the majority of all possible edges would already be included into networks at this threshold. For each examined threshold, we balance between the number of nodes included into networks and network density. Empirically, we find that PC threshold of 0.95 results in the most appropriate networks, i.e., in networks that include many nodes while still being sparse enough. Also, since each GSR signal consists of $n = 1,000$ samples (Section II-A), and since thus PC coefficient between two responses is computed by contrasting two 1,000-dimensional vectors, using a closed formula for statistical significance of PC at $n = 1,000$ [32], we find that threshold of 0.95 has a p-value of below 10^{-7}. Thus, we adopt 0.95 as the threshold for PC positive networks, including $\sim$30% of all possible edges into all 89 PC positive networks combined. (Not all networks necessarily have the same number of edges.) For a fair comparison, we adopt the same threshold for PC positive-negative networks.

However, the distribution of all possible PC scores and all possible MI scores is somewhat different. For example, for PC, $\sim$30% of all possible edges are included into networks at threshold of 0.95, while for MI, $\sim$75% of all edges are included into networks at the same threshold. Thus, for a fair comparison, instead of selecting the same threshold across the two different similarity measures, we adopt MI threshold of 0.975, as it also results in $\sim$30% of all possible edges being included into MI positive networks.

Henceforth, we report results only for PC positive and PC positive-negative networks at threshold of 0.95 and for MI positive networks at threshold of 0.975.

B. Topological properties of constructed networks

We study six topological properties of the networks of different types: the number of nodes, number of edges, size of the largest connected component, maximum diameter, average clustering coefficient, and average degree (Section II-C). Since topological trends are very similar for PC positive and PC positive-negative networks, except that PC positive-networks are slightly denser, in this section, we illustrate results for PC positive networks only. But all results hold for PC positive-negative networks as well. Then, we contrast results for PC positive networks against results for MI positive networks.

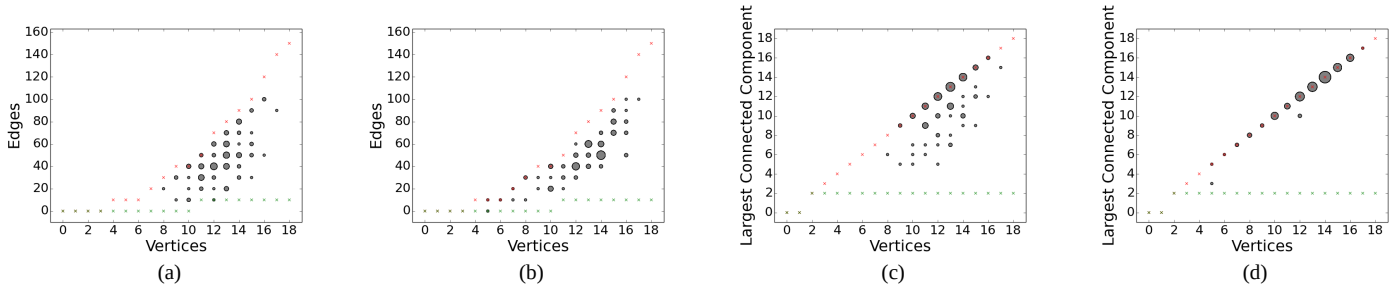
We find that in general there is a notable variability in topology of networks constructed in different ways, which could affect network-based clustering of images (Section III-C). Namely, PC networks tend to have more nodes than MI positive networks (Fig. 2), while MI positive networks tend to be denser than PC networks (Fig. 2). Also, with respect to the number of edges, in PC networks, different networks of a given size (in terms of the number of nodes) have very different numbers of edges (Fig. 2 (a)). On the other hand, in MI positive networks, different networks of a given size tend to agree more on their numbers of edges (Fig. 2 (b)). Similar trends are observed with respect to the average degree (results not shown). Further, largest connected components of PC networks of a given size tend to be smaller than components of MI networks of the same size (Fig. 2 (c) and (d)). That is, PC networks typically contain multiple connected components (Fig. 2 (c)), while in almost each MI network, all of the nodes are contained within the network's largest (and thus only) connected component (Fig. 2 (d)).

Networks of different types have similar trends with respect to their diameters and average clustering coefficients. Diameters of all networks are relatively small and their clustering coefficients are relatively high. This is encouraging, as the observed behavior is typical for many real-world networks [3].

C. Network-based clustering of images

1) ***The effect of network construction and clustering parameters:*** With all combinations of different network construction and clustering parameters, we produce a total of 10,235 image partitions (Section II-D). That is, for each combination of network type (PC positive, PC positive-negative, and MI positive networks), clustering method (HIE and KM), distance measure (seven network-based measures and one non-network-based measure), and linkage method (single, complete, average, and weighted; only for HIE), we vary the desired number of clusters, k, from 1 to 89 (Section II-D). Henceforth, we only focus on the *best partition* over all 89 ks, i.e., the partition that is the most similar to the AV "ground truth" partition in terms of Adjusted for chance Information Distance (AID; Section II-D4). There are 92 and 23 such partitions for HIE and KM, respectively. Of these, we focus only on partitions that are *statistically significantly* similar to the AV partition (p-value ≤ 0.01; Section II-D4). There are 16 and nine such partitions for HIE and KM, respectively.

Given the "statistically significant" partitions, we evaluate the effect of the different parameters (network type, distance measure, and, for HIE, linkage) on the partition quality. For this, we ask, for each of the parameters, whether there is a



(a) (b) (c) (d)

Fig. 2: Illustration of topological properties of constructed networks. We plot the distribution of edge counts over (a) PC positive and (b) MI positive networks with the given number of nodes (or vertices) (as shown on x-axis). Also, we plot the distribution of the sizes of the largest connected components (in terms of the number of nodes) over (c) PC positive and (d) MI positive networks with the given number of nodes (as shown on x-axis). In each panel, the larger the number of networks with given properties, the larger the size of the corresponding circle. Also, in each panel, green and red crosses correspond to theoretical minimum and maximum values of given properties, respectively. For example, in a network with n non-isolated nodes, there has to be at least $\lceil \frac{n}{2} \rceil$ edges (colored in green in panels (a) and (b)) and there can be at most $\binom{n}{2}$ edges (colored in red in panels (a) and (b)). Or, in a network with n non-isolated nodes where $n \geq 2$, the largest connected component has to have at least 2 nodes (colored in green in panels (c) and (d)) and it can have at most n nodes (colored in red in panels (c) and (d)).

single value of this parameter that "works" (i.e., produces "statistically significant" partitions) for all other parameter combinations. For example, we ask whether any of the three network types works for all combinations of distance measures and linkages (for HIE). If the answer is yes, then we can safely discard all other values of that parameter without sacrificing quality of the results. However, if the answer is no, it would mean that the results are highly dependent on the parameter choice. In this case, we ask which parameter value is the "best" (i.e., works for the most combinations of the other parameters).

HIE. We first determine relative performance of different network types, while fixing the choice of distance measure and linkage. PC outperforms MI (Table I). Specifically, for the seven network-based measures, out of 28 possible distance measure-linkage pairs, PC positive-negative, PC positive, and MI positive networks produce "statistically significant" partitions in six, three, and four cases, respectively. In addition, for the NON distance measure, only PC results in "statistically significant" partitions, while MI does not. Importantly, none of the network types works for all seven network-based distance measures, even when we ignore linkage; here, the best network type are PC positive-negative networks, which work for four out of the seven measures. Interestingly, only PC positive-negative networks work for RGF-distance and Pearson correlation of degree distribution. On the other hand, all three network types work for three out of four linkages (complete, average, and weighted), when we ignore distance measure; but only PC positive-negative networks work for single linkage.

Next, we determine relative performance of different distance measures, while fixing the choice of network type and linkage. Common edges and the difference of average clustering coefficients are superior, producing "statistically significant" partitions for four out of twelve possible network type-linkage pairs. Interestingly, contrary to our expectation that highly constraining measures of network similarity would work the best (Section II-D1), GDD-agreement and Pearson correlation of degree distributions do not produce any "statistically significant" partitions. That is, topologically "simpler" measures can give better results that these highly constraining measures. Importantly, no distance measure works for all three network types, even when ignoring linkage, nor for all four linkages, even when ignoring network type. In particular, each distance measure produces "statistically significant" partitions for either PC- or MI-based similarities, the only exception being common edges, which works for both PC positive and MI positive networks. Note that only the network-based measures produce "statistically significant" partitions for MI, while the NON distance measure fails for MI.

Finally, we determine relative performance of different linkages, while fixing the choice of network type and distance measure. Out of 21 possible network type-distance measure combinations for network-based measures and two possible combinations for the NON distance, average, complete, weighted, and single linkage work for six, five, three, and two of them, respectively. Each of average, complete, average, and weighted linkage works for all three network types when we ignore distance measure. On the other hand, no linkage works with all eight distance measures even when we ignore network type (or similarity measure for the NON distance). The superior linkage in this sense is average, which works for five out of the eight measures. And while single linkage works for only two distance measures, it is the only linkage that works for Pearson correlation of degree distributions.

KM. Here, we study the effect of two parameters - network type and distance measure. Similarly as for HIE, PC generally works better than MI, with PC positive-negative networks demonstrating the best results. Regarding the choice of distance measure, the common edges measure works for all three network types, and also, the NON distance works for both response similarity measures. On the other hand, GDD-agreement, RGF-distance, and Pearson correlation of clustering spectra, the topologically constraining measures, again produce unexpected results (Section II-D1): they never lead to "statistically significant" partitions.

Summary. Our analysis suggests that the choice of network construction and clustering parameters affects the resulting partitions. Importantly, in general, there is no single choice of any of the parameters that works for all combinations of the other parameters. This is especially true for HIE. This

implies that we still have to consider all statistically significant partitions and choose the best according to the desired criteria (Section II-D4). For the network research community, our study suggests the importance of considering various parameters instead of focusing on a single combination, unless there is a proper justification for that. Yet, our finding is not necessarily alarming, as even though different combinations of the parameters result in different "statistically significant" partitions, the majority of the partitions are actually statistically significantly similar to each other (with respect to AID).

The fact that we are able to construct statistically significant partitions using a network-based approach implies that differences in physiological response patterns of subjects to various images captured by our approach are meaningful with respect to the AV "ground truth" partition. Also, the network-based partitions tend to fit the AV partition better than the NON-distance-based partitions (Table II), indicating that network analysis indeed can improve interpretation of the data.

We note that none of the "statistically significant" partitions ideally matches the AV partition, as their AID scores are above zero (Table II). A possible explanation could be the fact that we rely on the GSR signal, which is traditionally considered to be sensitive to changes in arousal [11], while the AV partition is based both on arousal and valence levels [10]. Another explanation could be the noisiness of the AV "ground truth" partition (Section II-A). And because of this noisiness in the "ground truth" data, we aim to answer whether our network-based partitions improve upon the AV partition, as follows.

	Linkage method			
Distance measure	Single	Complete	Average	Weighted
NON distance	-	PC	PC	PC
Common edges	-	PC_p,MI_p	PC_p	MI_p
GDD-agreement	-	-	-	-
RGF-distance	-	-	PC_{pn}	PC_{pn}
Pearson correlation of degree distributions	PC_{pn}	-	-	-
Pearson correlation of clustering spectra	-	-	-	-
Difference of average diameters	-	MI_p	MI_p	-
Difference of average clustering coefficients	PC_{pn}	PC_{pn}	PC_p,PC_{pn}	-

TABLE I: Network types, or, for NON distance, network similarity measures, that result in "statistically significant" partitions, out of three possible network types (PC positive ("PC_p"), PC positive-negative ("PC_{pn}"), and MI positive ("MI_p") networks) or out of two response similarity measures ("PC" and "MI"), for different combinations of fixed clustering distance measures and linkage methods. Clearly, none of the combinations works for all three network types. Similar results hold when fixing network types and linkage methods: none of their combinations works for all distance measures. Also, similar results hold when fixing network types and distance measures: none of their combinations works for all linkages.

*2) **Semantic analysis of image partitions**:* To further evaluate the quality of the "statistically significant" partitions and their potential improvement over the AV partition, we measure their semantic meaningfulness using LSA (Section II-D4). To be able to compare HIE and KM fairly, we remove redundant linkages for HIE: if the same combination of network type and distance measure produces "statistically significant" partitions for multiple linkages, we keep only the linkage with the lowest AID score. This leaves us with 17 "statistically significant" partitions: eight for HIE and nine for KM. And since the AV "ground truth" partition is also a partition of the image set, we measure its semantic meaningfulness using LSA as well.

Then, we ask whether our network-based partitions outperform in terms of LSA: 1) the AV partition and 2) the NON-based partitions. If so, that would mean that: 1) even though our partitions do not perfectly match the AV partition (but are still statistically significantly similar to it), they are more meaningful (according to LSA), and 2) they are more meaningful than the non-network based analysis of the same physiological data that we employed in our study. That is, this would further confirm the validity of our network analysis strategy in the context of affective physiological data.

Indeed, this is exactly what we observe (Table II). While the LSA p-value for the AV partition is 0.261, which means that it is not statistically significant, five of our partitions (four for HIE and one for KM) are statistically significantly "meaningful" in terms of LSA. Importantly, all of them are network-based. This confirms that network analysis can improve interpretation of physiological data.

IV. Conclusions

We use a network-based approach to study affective physiological data. Namely, we model different images as networks and group together images with similar network topologies. We perform a comprehensive and systematic analysis of the effect of different network construction and clustering parameters, concluding that the choice of each parameter can affect the results. For network inference and clustering communities, this highlights the importance of considering various strategies and their parameters. Nonetheless, we show that via network analysis we can construct image partitions that are statistically significantly similar to the "ground truth" partition, while at the same time they improve it. Importantly, we show that such a result cannot be obtained via a non-network-based analysis of the same data. Thus, viewing affective physiological data through a network lens can improve analysis of the data.

We introduce a framework for systematic network analysis of human physiological responses. There are several future extensions of our research. While we construct image networks modeling similarities between responses of different individuals to a given image, it is also possible to construct subject networks, modeling similarities between responses of the given subject to different images, which could lead to complementary insights. While we use a threshold-based approach for network construction, alternative approaches, e.g., the extraction of a minimum spanning tree [33] or the network deconvolution model [34], can be used. While we focus on GSR signals, our framework can be applied to other signals, e.g., ECG or EMG, which would allow to study relationships between different physiological channels. This could lead to the ultimate goal of understanding how different emotions are manifested in physiological signals both within and across individuals.

V. Acknowledgements

Funding: NSF CCF-1319469 and EAGER CCF-1243295 grants to the last author, and NSF HCC-0834847 and DRL-1235958 grants to the second author.

References

[1] D. Watts and S. Strogatz, "The small world problem," *Collective Dynamics of Small-World Networks*, vol. 393, pp. 440–442, 1998.

[2] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," *Cancer informatics*, vol. 6, p. 257, 2008.

GSR signal similarity measure/Network type	Clustering method	Distance measure	AID score	LSA p-value
PC positive	Hierarchical, average linkage	Common edges	0.887	0.023
PC, non-network-based	Hierarchical, complete linkage	NON distance	0.923	0.232
MI positive	**Hierarchical, weighted linkage**	**Common edges**	**0.924**	**0.001**
PC positive-negative	**Hierarchical, single linkage**	**Difference of average clustering coefficients**	**0.929**	**1.72E-10**
PC positive-negative	**Hierarchical, single linkage**	**Pearson correlation of degree distributions**	**0.934**	**9.58E-07**
PC positive-negative	Hierarchical, average linkage	RGF-distance	0.940	0.013
MI positive	**Hierarchical, average linkage**	**Difference of average diameters**	**0.942**	**1.04E-05**
PC positive	Hierarchical, average linkage	Difference of average clustering coefficients	0.944	0.203
PC, non-network-based	K-medoids	NON distance	0.902	0.123
PC positive	**K-medoids**	**Difference of average clustering coefficients**	**0.910**	**0.008**
PC positive-negative	K-medoids	Common edges	0.929	0.051
PC positive-negative	K-medoids	Pearson correlation of degree distributions	0.929	0.246
PC positive	K-medoids	Common edges	0.932	0.065
MI positive	K-medoids	Common edges	0.933	0.177
PC positive-negative	K-medoids	Difference of average clustering coefficients	0.935	0.416
MI, non-network-based	K-medoids	NON distance	0.939	0.403
PC positive-negative	K-medoids	Difference of average diameters	0.941	0.526

TABLE II: Quality of the "statistically significant" image partitions with respect to AID score and LSA p-value. The lower the AID score, the more similar the given partition to the AV partition. All AID scores are statistically significant at p-value threshold of 0.01. The lower the LSA p-value, the more semantically meaningful the partition. For each clustering method (HIE and KM), partitions are listed in the increasing order of their AID scores. Partitions with LSA p-value lower than 0.01 are shown in bold.

[3] M. Newman, *Networks: an introduction*. Oxford University Press, 2009.

[4] C. Izard, "The many meanings/aspects of emotion: Definitions, functions, activation, and regulation," *Emotion Review*, vol. 2, no. 4, pp. 363–370, 2010.

[5] R. Picard, *Affective Computing*. Cambridge, Mass: MIT Press, 1997.

[6] R. A. Calvo and S. K. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[7] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[9] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, *Emotion Assessment: Arousal Evaluation Using EEGs and Peripheral Physiological Signals*, 2006, pp. 530–537.

[10] O. Alzoubi, S. D'Mello, and R. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *Affective Computing*, pp. 298–310, 2012.

[11] J. Larsen, G. Berntson, K. Poehlmann, T. Ito, and J. Cacioppo, *The psychophysiology of emotion*, 3rd ed. New York, NY: Guilford, 2008, pp. 180–195.

[12] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] B. Herbelin, P. Benzaki, F. Riquier, O. Renault, and D. Thalmann, "Using physiological measures for emotional assessment: A computer-aided tool for cognitive and behavioral therapy," *International Journal in Disability and Human Development*, vol. 4, no. 4, pp. 269–277, 2004.

[14] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, "Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback," *International Journal of Human-Computer Interaction*, vol. 25, no. 6, pp. 506–529, 2009.

[15] M. van der Zwaag, J. Janssen, and J. Westerink, "Directing physiology and mood through music: Validation of an affective music player," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 57–68, 2012.

[16] K. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[17] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," *Technical Report A-8*, 2008.

[18] J. Russell, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, pp. 145–172, 2003.

[19] A. Dionísio, R. Menezes, and D. A. Mendes, "Mutual information: a dependence measure for nonlinear time series," *EconWPA Econometrics series*, no. 0311003, 2003.

[20] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 9999, pp. 2837–2854, 2010.

[21] T. Milenković, J. Lai, and N. Pržulj, "GraphCrunch: a tool for large network analyses," *BMC bioinformatics*, vol. 9, no. 1, p. 70, 2008.

[22] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[23] R. W. Solava, R. P. Michaels, and T. Milenković, "Graphlet-based edge clustering reveals pathogen-interacting proteins," *Bioinformatics*, vol. 28, no. 18, pp. i480–i486, 2012.

[24] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.

[25] O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj, "GraphCrunch 2: Software tool for network modeling, alignment and clustering," *BMC bioinformatics*, vol. 12, no. 1, p. 24, 2011.

[26] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[27] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, pp. e177–e183, 2007.

[28] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj, "Optimal network alignment with graphlet degree vectors," *Cancer informatics*, vol. 9, p. 121, 2010.

[29] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," *Statistical Data Analysis Based on the L1 Norm*, 1987.

[30] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1073–1080.

[31] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[32] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2003.

[33] R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 11, no. 1, pp. 193–197, 1999.

[34] S. Feizi, D. Marbach, M. Médard, and M. Kellis, "Network deconvolution as a general method to distinguish direct dependencies in networks," *Nature biotechnology*, 2013.