Journal of Complex Networks (2014) **2**, 614–636
doi:10.1093/comnet/cnu032
Advance Access publication on 10 August 2014

Network analysis improves interpretation of affective physiological data[†]

Yuriy Hulovatyy
Department of Computer Science and Engineering, ECK Institute for Global Health, and Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN 46556, USA

Sidney D'Mello
Department of Computer Science and Engineering and Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, USA

Rafael A. Calvo
School of Electrical and Information Engineering, University of Sydney, Darlington, Sydney, NSW 2006, Australia

and

Tijana Milenković[‡]
Department of Computer Science and Engineering, ECK Institute for Global Health, and Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN 46556, USA
[‡]Corresponding author. E-mail: tmilenko@nd.edu

Edited by: Ernesto Estrada

[Received on 22 February 2014; accepted on 26 June 2014]

Understanding how human physiological responses to stimuli vary across individuals is critical for the fields of Affective Psychophysiology and Affective Computing. We approach this problem via network analysis. By analysing 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 statistically 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 yield insights that go beyond 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). Non-network-based approaches do not yield any such insights. Thus, network analysis of affective physiological data appears to improve interpretation of the data. We conclude by analysing in-depth a representative network-based image partition and discussing practical applications of the corresponding results.

[†]This work is an extended version of our initial conference publication [1].

Keywords: pattern clustering; physiology; GSR; affective computing; affective physiological data interpretation; arousal/valence-based ground truth partition; galvanic skin responses; network analysis; network-based image partitions; network-based partitions; affective physiological data; complex networks; latent semantic analysis.

1. Introduction

Networks (or graphs) allow for studying complex processes that emerge from the collective behaviour of interconnected elements. Thus, networks can model real phenomena in many domains, e.g. social, technological or biological systems [2–5]. We focus on network modelling of affective (or emotional) 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. While networks have already been used to evaluate emotional responses (e.g. [6]), to our knowledge, we are the first to apply them to study affective physiological data. As we will show, network analysis improves interpretation of the data compared with an alternative non-network-based approach.

1.1 *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 [7]. Emotions are constructs (or conceptual entities) that cannot be directly measured, but must be inferred from measurable signals like physiology. Thus, understanding the link between emotions and physiology has been an important endeavour in the field of Affective Psychophysiology for more than a century.

There is also an engineering side to complement the scientific endeavour of identifying the physiological correlates of affect. The field of Affective Computing, a subfield of Human–Computer Interaction, aims to build intelligent systems that respond to user emotions much like an actual human would [8]. 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. (e.g. [9,10]). Physiological-based approaches for affect detection are attractive (e.g. [11–13]), as 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.

1.2 *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 [14] for a review). While it was once claimed that unique discrete emotions (e.g. fear, anger) are accompanied by distinct physiological patterns [15], meta-analyses and other syntheses of the literature have failed to conclusively support this claim [14].

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. [12,16,17]). This is done by building *person-dependent* models that are carefully calibrated to each individual. However, despite some advances [11,18], what is needed are *person-independent* models that generalize to new individuals. The few efforts along this front

have produced mixed results [19]. For example, emotion recognition accuracy of person-dependent and person-independent models from three physiological signals (ECG, EMG and GSR) were compared [13], 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 with person-independent modelling.

1.3 *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. Modelling 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 modelling 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 analyse 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. Moreover, we demonstrate that network analysis of such data improves interpretation of the data compared with an alternative, non-network-based approach. Finally, we analyse in detail a representative network-based image partition and discuss practical applications of its results.

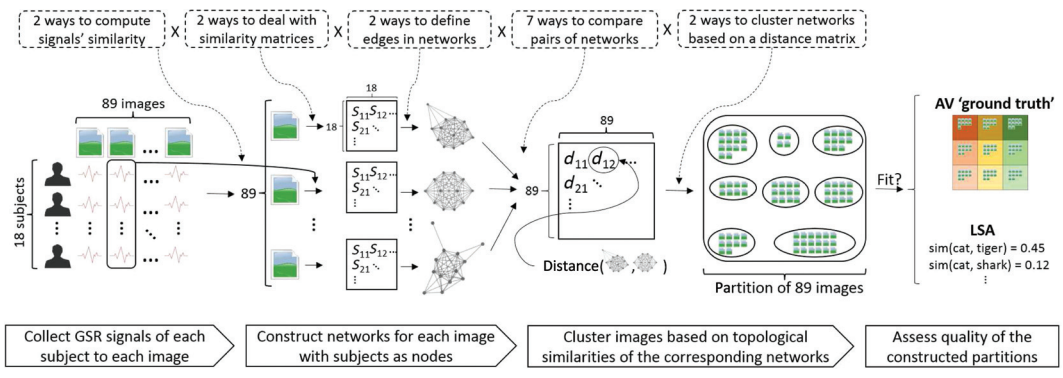

FIG. 1. Overview of our four-step study. Each of the four steps is discussed in one of the four subsections of Section 2.

2. Methods

Our study consists of four major steps (Fig. 1). First, we obtain affective physiological data (Section 2.1). Secondly, we construct networks from these data using a series of network inference strategies (Section 2.2). Thirdly, we use various clustering methods and combinations of their parameters to partition the set of networks based on their topological similarities (Section 2.3). Fourthly, we conduct a thorough evaluation of the partitions produced by the different clustering strategies, and verify both statistical and practical significance of our results (Section 2.4).

2.1 *Data collection*

During the experiment, 18 human subjects were presented with 89 emotionally charged images from the International Affective Picture System (IAPS) [20–22]. The IAPS is a collection of over a 1000 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 [23].

The 89 images were selected to cover a 3×3 arousal–valence (arousal: low, medium, high; valence: negative, neutral, positive) normative rating space with 10 images per each of the nine classes (one class had only nine images due to error) (Fig. 2(a)). For example, the images in the positive valence and high-arousal class were selected using the normative ratings above 6.0 for valence and above 5.5 for arousal (normative ratings ranged from 1 to 9). Selection of images proceeded in a three-step process. In Step 1, the images were categorized into the nine classes of the 3×3 valence–arousal space based on the normative IAPS ratings. In Step 2, the researchers then manually selected images from each class to ensure that there was sufficient variability in content and minimal overlap across classes. In Step 3, mean valence and arousal scores for the selected images in each class were analysed in order to ensure that they were sufficiently different. For reproducibility of our study, IAPS identifiers of the selected 89 images, along with their corresponding classes in the arousal–valence partition, are presented in Appendix A.

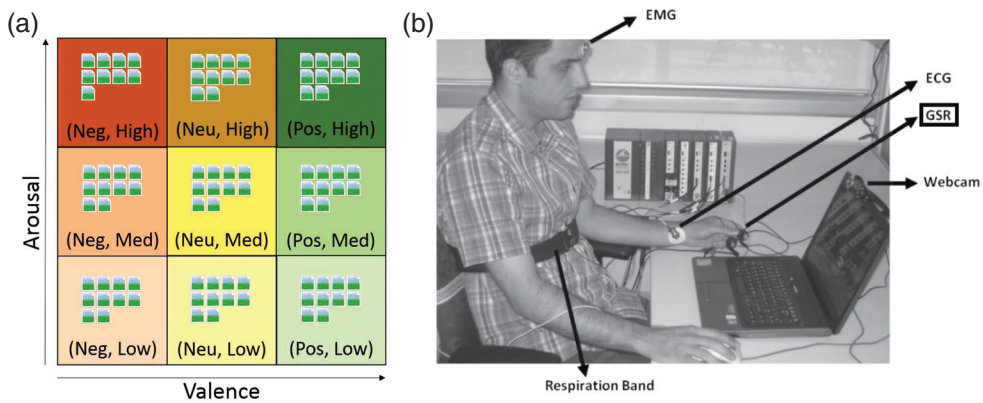

FIG. 2. Illustration of physiological data used in our study with respect to: (a) AV 'ground truth' partition of the image set and (b) experimental setup. In (a), for each image, its IAPS arousal normative score was categorized as either 'low', 'medium' or 'high', and its IAPS valence normative score was categorized as either 'negative', 'neutral' or 'positive'. Based on this 3×3 AV space, images were partitioned into nine classes.

Henceforth, we refer to this nine-class image partition as the *arousal/valence* (*AV*) *'ground truth' partition* (or simply as the *AV partition*). 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 is noisy, which is typical for any emotion-eliciting stimulus.

For physiological recording, participants were equipped with ECG, EMG, respiration and GSR sensors (Fig. 2(b)) [24]. The BIOPAC MP1501 system with AcqKnowledge software was used to acquire the physiological signals at a sampling rate of 1000 Hz for all channels. Two electrodes were placed on the wrists for collecting ECG. Two channels recorded EMG activity from two facial muscles regions (zygomatic and corrugator, respectively). GSR was recorded from the index and middle finger of the left hand, and a respiration band was strapped around the chest.

We exclusively focus on the GSR signal as an initial step to studying the viability and added-value of the network-based approach to analysing affective psychophysiological data. The GSR was used in lieu of the other signals due to the considerable research establishing a clear link between GSR and physiological arousal; this link is less clear for the other signals (see [14] for a review). 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. We note that even though here we focus exclusively on the GSR signal, our study is easily extensible to any of the recorded signals.

Each subject viewed each image for 10 s, with responses being recorded at a rate of 1000 Hz. There was a 6-s break between image presentations in order to allow the signals to return to baseline values. A GSR signal of each subject to each image was obtained, resulting in $18 \times 89 = 1602$ signals. 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.

2.2 *Network construction*

We construct networks from the collected physiological response data as follows. 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' with respect to a given similarity measure. Thus, each network structurally captures how different subjects respond to the same image as an initial step in modelling inter-individual variability.

To form a network, we need to choose: (1) a similarity measure for comparing GSR signals of two nodes, (2) a method for dealing with (e.g. filtering in some way) the resulting pairwise node similarity matrix and (3) a strategy for defining from this matrix edges between the nodes (Fig. 1). To obtain representative results, we evaluate the effect of the choice of these network construction parameters on our results. Namely, we test two similarity measures (Section 2.2.1). For each of the measures, we use two different strategies for dealing with the resulting matrices (Section 2.2.2). Finally, for each matrix, we define edges in the network in two ways (Section 2.2.3).

2.2.1 *Signal similarity measures.* We measure GSR similarity by using: (1) Pearson correlation (PC) and (2) mutual information (MI). While PC measures linear relationships between signals, MI can

capture more complex (non-linear) relationships between signals by relying on an information-theoretic notion of similarity [25,26]. We normalize MI to take values between 0 and 1 [25].

2.2.2 *Dealing with signal similarity matrices.* For each image, we perform pairwise node comparison of the corresponding responses in order to obtain an 18×18 signal similarity matrix. Then, we either: (1) directly use this matrix to define edges (as described below) or (2) process this original matrix using the network deconvolution method [27] into a filtered matrix and then use the filtered matrix to define edges (as described below). The motivation for applying the network deconvolution method to the original matrix is as follows [27]. When constructing an *observed* network from an unfiltered matrix of pairwise node (i.e. subject) similarities (such as correlations), the resulting edges will likely include numerous indirect dependencies owing to transitive effects of the similarities. For example, if there is a strong dependency between subjects X and Y and between subjects Y and Z in the underlying *true* network, high similarities will likely also be visible between subjects X and Z in the observed network, thus inferring an edge from subjects X to node Z which might not actually exist in the true network. Moreover, even if this edge did exist in the true network, owing to additional indirect relationships between X and Z, the strength of this edge may be overestimated; consequently, this true edge could be wrongly included into the observed network over some other true edge with higher initial strength but no indirect relationships to boost its strength up. Hence, filtering the matrix of node similarities by distinguishing between the convolved direct and indirect contributions could ensure that the resulting observed network matches the true network better than the observed network resulting from the non-filtered matrix of node similarities. As such, matrix filtering could lead to more meaningful results, and this is exactly what we examine. We refer to networks directly constructed from the original signal similarity matrices as *unfiltered* networks, and we refer to networks constructed from the filtered matrices as *filtered* networks.

2.2.3 *Defining edges.* We define edges from a signal similarity matrix (unfiltered or filtered) as follows. We add an edge between two nodes if the similarity score between the corresponding subjects' responses to the given image is above some threshold. We refer to the resulting networks as *positive* networks. In addition, we add an edge between two nodes if the *absolute value* of the similarity score is above some threshold. We refer to the resulting networks as *positive–negative* networks. Hence, in PC positive–negative networks, both subjects whose responses are strongly positively correlated as well as subjects whose responses are strongly negatively correlated are linked. Note that since MI values are non-negative, positive–negative networks are constructed only with respect to PC.

To form a network (positive or positive–negative) from a signal similarity matrix (unfiltered or filtered), we need to specify an edge threshold value. We aim to choose this value in a way that keeps only statistically significant edges and provides a meaningful representation as well as interpretation of the data [1]. Namely, for each image, 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 [4].

We begin by focusing on PC positive networks (Section 2.2). 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, as discussed above. Empirically, we find that PC threshold of 0.95 (with p-value of below 10^{-7} [1]) results in the most appropriate networks, i.e. in networks that include many nodes while still being sparse enough. Thus, we adopt 0.95 as the threshold for PC positive networks, including ∼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. Since the distribution of all possible PC scores and all possible MI scores is somewhat different, for MI positive networks, we do not select the same threshold; instead, for a fair comparison, we adopt MI threshold of 0.975, as this threshold also results in ∼30% of all possible edges being included into MI positive networks.

2.2.4 *Bottom line*: *network types.* To summarize, we have six different types of networks: PC positive, PC positive–negative and MI positive, each with both an unfiltered and filtered version. For each of the six types, we generate 89 networks, one for each of the 89 images. Each network models similarity in responses between the 18 subjects and thus has 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.

2.3 *Clustering of networks*

After constructing the networks, we next ask whether networks corresponding to images of the same AV 'ground truth' class (Section 2.1) 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 2.1), 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) [28,29].) Instead, we consider each network as a separate 'atomic' object and perform 'data clustering' of a *set of networks* into groups of 'related' networks.

To cluster a set of objects (i.e. images), we need to define: (1) a measure of distance (or equivalently, similarity) between the objects, (2) a clustering method and (3) parameters of the method. We comprehensively test multiple network-based (as well as non-network-based) distance measures (Section 2.3.1), clustering methods (Section 2.3.2) and variations of the methods' parameters (Section 2.3.2).

2.3.1 *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 ($|E_1 \cap E_2|/|E_1 \cup E_2|$, where E_1 and E_2 are the two edge sets) [30]; (2) absolute *difference of* the networks' *average clustering coefficients* [31]; (3) absolute *difference of* their *average diameters* (the average diameter of the network is the average of shortest path lengths over all node pairs [31]); (4) *PC of* the networks' *degree distributions* [4,31]; (5) *PC of* their *clustering spectra* (the clustering spectrum of a network is the distribution of average clustering coefficients of nodes with a particular degree) [31]; (6) relative graphlet frequency distance (*RGF-distance*) (which compares frequencies of all three- to five-node subgraphs, or *graphlets*, in two

networks [32]) and (7) graphlet degree distribution agreement (*GDD-agreement*) (which generalizes the degree distribution into a spectrum of GDDs [33]). We use *GraphCrunch* for all comparisons [31,34].

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. For this purpose, we define an additional non-network-based measure of similarity between two images, called *NON-distance*, that is, based solely on the direct correlation of signal similarity matrices constructed in Section 2.2. For details, see our previous work [1].

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

2.3.2 *Clustering methods and their parameters.* To test how the choice of clustering method affects a partition quality, we use two clustering methods: (1) hierarchical clustering (HIE) and (2) k-medoids clustering (KM) [28]. Note that KM clustering is a modification of k-means clustering that requires using actual data points (in our case, networks) as cluster centres, rather than allowing centres to be non-data points, as k-means does. And since our clustering distance measures require centres to be data points, using k-means is inappropriate, or in other words, KM needs to be used. We test various parameters for the two clustering methods [1]: four linkages for HIE (*single*, *complete*, *average* and *weighted* [28]) and all possible values of the desired number of clusters, k, for both HIE and KM ($k = 1, 2, \ldots, 89$).

All combinations of the two signal similarity measures (PC and MI), two ways to deal with signal similarity matrices (using either the original unfiltered matrix or the filtered matrix), two ways to define edges (positive and positive–negative), eight distance measures (seven network-based measures and NON-distance) and two clustering methods (HIE and KM) lead to 96 different parameter combinations: 48 combinations for unfiltered networks and 48 combinations for filtered networks. Actually, since for unfiltered networks, NON-distance produces the same distance matrix for PC positive networks and PC positive–negative networks, and since it does so for both hierarchical and k-means clustering, the actual number of different combinations for unfiltered networks is 46 instead of 48. Thus, in total, over both unfiltered and filtered networks, there are $46 + 48 = 94$ different parameter combinations. For each of these combinations, we choose the 'best' partition over all combinations of clustering parameters (over all possible choices of k and linkage for HIE, and over all possible choices of k for KM). By 'best', we mean the most significant according to the criterion introduces in the following section.

2.4 *Evaluating partition quality*

Upon producing a partition of images, we evaluate the quality of the partition with respect to its overlap with the 'ground truth' knowledge about the images. Namely, we evaluate a partition: (1) by comparing it against the AV 'ground truth' partition and (2) according to its semantic meaning.

To compare two partitions, we use *Adjusted for chance Information Distance* (AID) measure [26,35]. This measure uses notions of entropy and MI to determine the similarity between two partitions from an information-theoretic perspective. It quantifies how much knowing one of the partitions reduces uncertainty about the other [26]. 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 [35]. 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 better, i.e. closer to the 'ground truth'. We determine the statistical significance of our AID score between two partitions empirically, as the percentage of 10^6 randomly generated partitions have the same or better AID scores than our actual AID score [1]. This percentage is our p-value. If an AID score between some partition and the AV 'ground truth' partition has a p-value below 0.01, we refer to that partition as 'statistically significant'. For further analysis, among all 94 best partitions (corresponding to 94 different parameter combinations), we focus only on those that are statistically significantly similar to the AV partition with respect to AID.

In addition to comparing our 'statistically significant' partitions with the AV 'ground truth' partition, we also assess them using latent semantic analysis (LSA) [36]. 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. For each pair of our images, we derive LSA similarities between high-level image names (e.g. kittens, garbage, spider) depicted in the images using the online LSA tools (http://lsa.colorado.edu/). That is, we obtain an additional 'ground truth' data set based purely on high-level semantic meanings of the images. We evaluate partition quality from the LSA perspective by comparing intra- and inter-cluster LSA similarities within a partition using the Wilcoxon rank-sum test [1]. We note that by using high-level labels to categorize each image, we rely on somewhat primitive keywords of images (e.g. dog) instead of more informative descriptions (e.g. dog stretching out in the lawn gazing at mail man). We do this because our primary goal is to capture primitive semantic influences on physiological early responding triggered by the 'gist' of the image rather than more complex cognitive appraisals of the information depicted in each image.

3. Results and discussion

We first discuss results for unfiltered networks (Section 3.1), followed by results for filtered networks (Section 3.2), in order to evaluate the effect of filtering on the quality of the results (Section 3.3). Also, we discuss results and implications of a representative network-based partition (Section 3.4).

3.1 *Results for unfiltered networks*

Here, we summarize topological properties of unfiltered networks (Section 3.1.1) as well as the quality of their partitions (Section 3.1.2). (More detailed results are also available [1].)

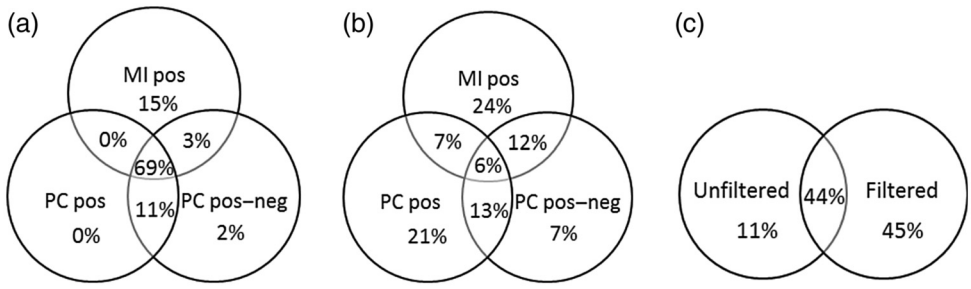

FIG. 3. Edge intersections between: (a) the three network types (PC positive, PC positive–negative and MI positive) for unfiltered networks, (b) the three network types for filtered networks and (c) unfiltered and filtered networks, ignoring the network type.

3.1.1 *Network topological trends.* We study several topological properties of networks of different types (Section 2.2.4), such as network size, size of the largest connected component, maximum diameter, average clustering coefficient and average degree [1]. While 69% of all edges in unfiltered networks are common to all three network types (Fig. 3(a)), there is still a notable variability in the topology of different networks types, which could affect network-based clustering of images. For example, even though topological properties are similar for PC positive and PC positive–negative networks (except that PC positive networks are slightly denser), PC networks have more (non-isolated) nodes than MI positive networks, while MI positive networks tend to be denser than PC networks. Moreover, PC networks typically contain multiple connected components, while in almost each MI network, all of the nodes are contained within the network's largest (and thus only) connected component. Networks of different types have similar trends only 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 to see in our data, as the observed behaviour is typical for many real-world networks [4].

3.1.2 *Quality of partitions.* After we construct networks, we use them to partition the image set (Section 2.3). We focus only on 'statistically significant' partitions (Section 2.4). There are 17 such partitions for unfiltered networks, out of 46 possible partitions (Table 1). When we evaluate the effect of different network construction and clustering parameters on the partition quality, we find that the choice of the parameters affects the resulting partitions. For example, we find that PC is generally better (i.e. it produces more 'statistically significant' partitions) than MI, with PC positive–negative networks demonstrating the best results. Regarding the choice of distance measure, common edges and the difference of average clustering coefficients are superior, contrary to our expectation that more constraining measures of network similarity, such as GDD-agreement and RGF-distance, would be the best. Importantly, we find that there is no single choice for 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 a desired criterion.

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-network-based partitions (Table 1), indicating that network analysis indeed can improve interpretation of the data.

To further evaluate the quality of the 'statistically significant' partitions, we measure their semantic meaningfulness using LSA (Section 2.4). Then, we ask whether our network-based partitions outperform in terms of LSA: (1) the AV partition and (2) the non-network-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 semantically 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 what we observe (Table 1). While the LSA p-value for the AV partition is 0.261 (and thus non-significant), five of our partitions (four for HIE and one for KM) are semantically meaningful in terms of LSA. Importantly, all of them are network based. Even though non-network-based partitions are 'statistically significant', none of them is semantically meaningful at the same time. This confirms that network analysis can improve interpretation of physiological data by capturing both AV 'ground truth' and semantics.

TABLE 1 *Quality in terms of AID p-values and LSA p-values of the 'statistically significant' image partitions resulting from clustering of unfiltered networks. Network construction and clustering parameters are listed under 'Partition description'. |C| denotes the number of clusters in the corresponding partition. The lower the AID p-value, the more similar the partition is to the AV 'ground truth' partition. 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 p-values corresponding to unfiltered networks. The three rightmost columns show AID p-values, LSA p-values, and |C| of the equivalent partitions resulting from the same network construction and clustering parameters but from filtered networks instead of unfiltered ones. These partitions of filtered networks are not necessarily 'statistically significant'. All AID and LSA p-values (corresponding to both unfiltered and filtered networks) lower than* 0.01 *are shown in bold*

Partition description				Unfiltered networks			Filtered networks		
#	Network type	Clustering method	Distance measure	AID *p*-value	LSA *p*-value	|C|	AID *p*-value	LSA *p*-value	|C|
1	PC positive	HIE	Common edges	**3.70E−05**	0.023	21	0.970	0.058	30
2	PC, non-network based	HIE	NON-distance	**1.15E−03**	0.232	3	0.951	0.013	16
3	MI positive	HIE	Common edges	**1.25E−03**	**1.00E−03**	21	0.072	**1.82E−15**	8
4	PC positive–negative	HIE	Diff. of avg. clustering coefficients	**2.16E−03**	**1.72E−10**	20	**7.02E−03**	0.031	29
5	PC positive–negative	HIE	PC of degree distributions	**3.25E−03**	**9.58E−07**	34	0.292	**1.01E−04**	13
6	PC positive–negative	HIE	RGF-distance	**5.68E−03**	0.013	30	0.031	0.027	58
7	MI positive	HIE	Diff. of avg. diameters	**6.40E−03**	**1.04E−05**	10	0.066	**8.20E−05**	6
8	PC positive	HIE	Diff. of avg. clustering coefficients	**7.47E−03**	0.203	11	**9.91E−03**	0.083	16
9	PC, non-network based	KM	NON-distance	**1.51E−04**	0.123	10	0.019	0.086	18
10	PC positive	KM	Diff. of avg. clustering coefficients	**3.40E−04**	**8.00E−03**	9	**5.95E−03**	0.823	11
11	PC positive–negative	KM	PC of degree distributions	**1.94E−03**	0.246	44	0.085	**3.00E−03**	69

continued.

TABLE 1 *continued.*

Partition description				Unfiltered networks			Filtered networks		
#	Network type	Clustering method	Distance measure	AID *p*-value	LSA *p*-value	|C|	AID *p*-value	LSA *p*-value	|C|
12	PC positive–negative	KM	Common edges	**2.00E−03**	0.051	22	**1.29E−03**	0.042	7
13	PC positive	KM	Common edges	**2.66E−03**	0.065	24	0.041	**5.30E−04**	3
14	MI positive	KM	Common edges	**2.93E−03**	0.177	30	**7.16E−03**	0.279	11
15	PC positive–negative	KM	Diff. of avg. clustering coefficients	**3.50E−03**	0.416	9	**7.12E−03**	0.075	12
16	MI, non-network based	KM	NON-distance	**4.69E−03**	0.403	16	**4.02E−03**	0.815	20
17	PC positive–negative	KM	Diff. of avg. diameters	**5.53E−03**	0.526	10	**5.76E−03**	0.058	26

3.2 *Results for filtered networks*

Next, we show topological (Section 3.2.1) and partition quality (Section 3.2.2) results for filtered networks.

3.2.1 *Network topological trends.* We find that the topology of filtered networks is extremely sensitive to the choice of network construction strategy: only 6% of all edges in filtered networks (compared with 69% for unfiltered networks) are shared among the three network types (Fig. 3(b)). Importantly, even within a single network type, there is a notable variability in most of the analysed network topological properties, which could have drastic effects on network-based clustering.

3.2.2 *Quality of partitions.* There are 21 'statistically significant' partitions of filtered networks, out of 48 possible partitions (Table 2). When we evaluate the effect of different network construction and clustering parameters on the partition quality, just as for unfiltered networks, we again find that the choice of the parameters affects the resulting partitions. But unlike for unfiltered networks, we find that more topologically constraining network distance measures (Section 2.3.1) are now superior, as we would expect: RGF-distance results in 'statistically significant' partitions for all three network types in case of HIE, and GDD-agreement does the same for all three network types in case of KM.

Although it is possible to obtain 'statistically significant' partitions from filtered networks using the NON-distance measure, importantly, AID scores of these partitions are worse than those of network-based partitions (Table 2). That is, the network-based partitions again tend to fit the AV partition better than the non-network-based partitions.

When we evaluate the 'statistically significant' partitions in terms of LSA, we find that they are more similar than AV 'ground truth' partition (Table 2): three of the partitions are semantically meaningful with p-value of 0.01 (recall that the LSA p-value for the AV 'ground truth' partition is 0.261 and is thus non-significant). Importantly, just as for unfiltered networks, none of the non-network-based partitions for filtered networks is semantically meaningful at the same time. However, unlike for unfiltered networks, we note that GDD-agreement, which we would expect to be superior to all other distance measures, is indeed superior for filtered networks, as it is the only measure that yields to semantically meaningful partitions for both clustering methods (HIE and KM) as well as for both signal similarity measures (PC and MI) (Table 2).

In summary, just as for unfiltered networks, clustering of filtered networks produces partitions that are statistically significantly similar to the AV 'ground truth' partition while fitting more closely to LSA than the AV 'ground truth' partition. Importantly, network-based partitions again lead to better results than non-network-based partitions, further confirming that network analysis can improve the interpretation of physiological data.

3.3 *Comparison of results for unfiltered and filtered networks*

Here, we contrast topological properties between unfiltered and filtered networks (Section 3.3.1). Then, we ask which of the two should be used from a practical point of view (Section 3.3.2).

3.3.1 *Network topological trends.* Even though we designed our study so that all network construction strategies result in the same average network density (Section 2.2), unfiltered and filtered networks

TABLE 2 *Quality in terms of AID p-values and LSA p-values of the 'statistically significant' image partitions resulting from clustering of filtered networks. Network construction and clustering parameters are listed under 'Partition description'. |C| denotes the number of clusters in the corresponding partition. The lower the AID p-value, the more similar the partition is to the AV 'ground truth' partition. 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 p-values corresponding to filtered networks. The three rightmost columns show AID p-values, LSA p-values and |C| of the equivalent partitions resulting from the same network construction and clustering parameters but from unfiltered networks instead of filtered ones. These partitions of unfiltered networks are not necessarily 'statistically significant'. All AID and LSA p-values (corresponding to both filtered and unfiltered networks) lower than* 0.01 *are shown in bold*

Partition description				Filtered networks			Unfiltered networks		
#	Network type	Clustering method	Distance measure	AID *p*-value	LSA *p*-value	$\|C\|$	AID *p*-value	LSA *p*-value	$\|C\|$
1	MI positive	HIE	GDD-agreement	**1.80E−04**	0.067	29	0.137	**5.00E−03**	21
2	PC positive	HIE	RGF-distance	**2.99E−04**	0.680	16	0.034	0.034	30
3	PC positive	HIE	PC of clustering spectra	**7.46E−04**	0.310	8	0.021	0.290	12
4	MI positive	HIE	RGF-distance	**1.41E−03**	0.845	25	0.030	0.517	57
5	PC positive–negative	HIE	Diff. of avg. clustering coefficients	**4.81E−03**	0.029	26	**7.47E−03**	0.203	17
6	PC positive–negative	HIE	GDD-agreement	**5.68E−03**	**1.90E−09**	23	0.228	0.090	29
7	PC positive	HIE	PC of degree distributions	**6.15E−03**	0.908	10	0.068	0.014	25
8	PC positive	HIE	Diff. of avg. clustering coefficients	**6.82E−03**	0.214	11	0.054	0.835	36
9	PC, non-network based	HIE	NON-distance	**8.50E−03**	0.017	7	**2.24E−03**	0.795	18
10	PC positive–negative	HIE	Diff. of avg. diameters	**9.27E−03**	0.041	30	0.155	0.760	12
11	PC positive–negative	HIE	RGF-distance	**9.46E−03**	**1.00E−03**	18	**5.68E−03**	0.309	12

12	PC positive–negative	KM	GDD-agreement	**1.12E−04**	0.684	16	0.026	0.721	26
13	PC positive	KM	GDD-agreement	**6.99E−04**	0.114	18	0.021	0.324	10
14	PC positive–negative	KM	Common edges	**1.29E−03**	0.042	7	**2.00E−03**	0.051	22
15	MI, non-network based	KM	NON-distance	**4.02E−03**	0.815	20	**4.69E−03**	0.403	16
16	PC positive	KM	PC of clustering spectra	**4.51E−03**	0.159	19	0.038	0.664	61
17	MI positive	KM	GDD-agreement	**5.46E−03**	**1.00E−03**	21	0.160	**1.02E−07**	2
18	PC positive–negative	KM	Diff. of avg. diameters	**5.76E−03**	0.058	26	**5.53E−03**	0.526	10
19	PC positive	KM	Diff. of avg. clustering coefficients	**5.95E−03**	0.823	11	**3.40E−04**	**8.00E−03**	9
20	PC positive–negative	KM	Diff. of avg. clustering coefficients	**7.12E−03**	0.075	12	**3.50E−03**	0.416	9
21	MI positive	KM	Common edges	**7.16E−03**	0.279	11	**2.93E−03**	0.177	30

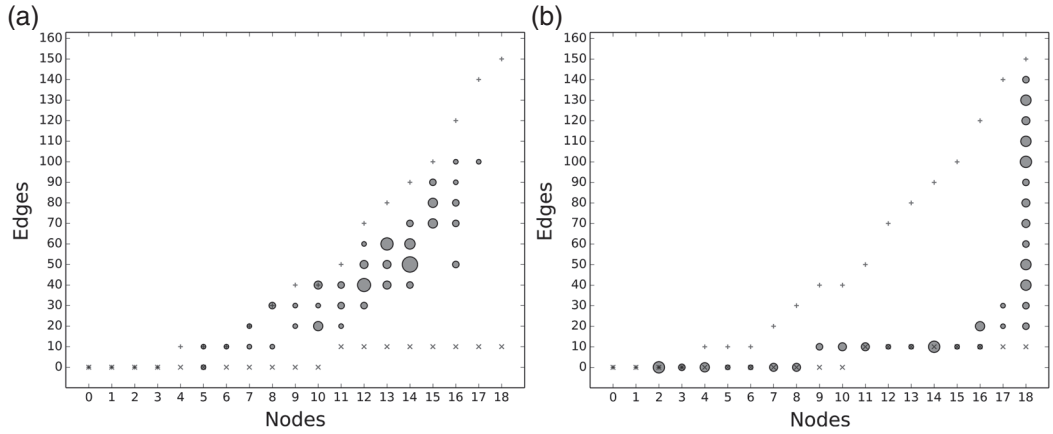

FIG. 4. Illustration of differences in topological properties of unfiltered and filtered networks. We plot the distribution of edge counts over MI positive networks with the given number of nodes (as shown on *x*-axis), constructed from: (a) unfiltered and (b) filtered signal similarity matrices. The larger the number of networks with the given property (i.e. the given number of nodes and edges), the larger the size of the corresponding circle. '×' and '+' marks in the panels correspond to theoretical minimum and maximum values of the illustrated properties, respectively.

differ in their topology. To start with, while the three network types share the majority of all edges in unfiltered networks (Fig. 3(a)), they share only a small portion of all edges in filtered networks (Fig. 3(b)). Also, when we ignore the network types and look at all edges in both unfiltered and filtered networks, only 44% of the edges are common to the two (Fig. 3(c)).

Going beyond simply counting the edges in the intersection between different network types, in general, unfiltered and filtered networks also have different topological properties. Namely, they differ in terms of network sizes (Fig. 3). Also, filtered networks have smaller clustering coefficients (or, intuitively, fewer triangles) than unfiltered networks. This could be because the network deconvolution filtering method is likely to decrease the effect of 'spurious' pairwise node similarities [27], and thus, potential edges contributing to indirect paths (such as the indirect path between two nodes in a triangle) could be weighted lower and consequently removed from the network.

3.3.2 *Unfiltered vs. filtered networks*: *which one to use from practical point of view?* Clearly, the network deconvolution method alters the topology of resulting filtered networks compared the topology of unfiltered networks. Still, both unfiltered and filtered networks produce statistically significant and semantically meaningful results (Tables 1 and 2, respectively). Yet, results that are significant and meaningful for unfiltered networks are not necessarily significant and meaningful for filtered networks (Table 1), and vice versa (Table 2). Thus, the question is which one to use: unfiltered or filtered networks? To answer this, we first contrast the two by measuring which one shows significant correlation of their network-based image distances with 'ground truth' image distances, before partitioning the images based on their network-based distances (this allows for a direct comparison of unfiltered and filtered networks without being confounded by clustering). Then, we contrast unfiltered and filtered networks with respect to the quality of their partitions produced based on the network-based image distances.

Correlation with 'ground truth' image distances. We directly correlate each of (non-)network-based image distance matrices with each of four 'ground truth' image distance matrices: (1) arousal (comparing IAPS normative arousal scores of the images), (2) valence (comparing IAPS normative valence scores of the images), (3) arousal–valence (comparing the combination of the previous two scores; Fig. 2(a)) and (4) LSA (comparing semantic similarity of the images). We do this via Mantel test, which is convenient for computing the significance of correlation between two distance matrices of related objects after adjusting for item-dependence [37].

We find that NON-distance, the non-network-based approach, captures (i.e. significantly correlates with) arousal, but nothing else. And capturing arousal is expected for GSR as repeatedly documented in the literature (e.g. [14,38]). Thus, the non-network-based approach does not seem to uncover any interesting (i.e. unexpected) signal from the data. Network-based approaches, on the other hand, capture all four types of the 'ground truth' image distance matrices. That is, network-based approaches go beyond simply capturing the expected arousal—they also significantly correlate with distance matrices of valence, arousal–valence and semantics (LSA). Hence, we again confirm that network analysis of the data has important practical applications, as it can capture interesting results that can easily be missed by non-network-based analysis.

Going back to determining which one is more efficient, unfiltered or filtered networks, we find that the latter outperform (in terms of the significance of correlations) the former for three out of the four 'ground truth' image distance matrices: arousal, valence and LSA. While both types of networks capture (i.e. significantly correlate with) arousal and LSA, only unfiltered networks capture arousal–valence, and only filtered networks capture valence. This suggests that perhaps the most defensible approach is to consider both filtered and unfiltered networks, though they may be added advantages to filtered networks as discussed below.

Quality of partitions. Unfiltered and filtered networks show comparable performance in terms of partition quality (Tables 1 and 2), although filtering leads to a slightly larger number of statistically significant partitions (21 for filtered networks vs. 17 for unfiltered networks) as well as semantically meaningful partitions (7 for filtered networks vs. 5 for unfiltered networks). However, the partition with the lowest AID score was obtained for a partition of unfiltered networks (partition 1 in Table 1).

We find that topologically more constraining network distance measures, such as GDD-agreement and RGF-distance, which we would expect to be superior, are indeed superior for filtered networks. For unfiltered networks, only one 'statistically significant' partition is obtained by using these measures, while there are eight such partitions for filtered networks (three of which are also semantically meaningful). Moreover, for filtered networks, both the partition with the lowest AID score for HIE, as well as the partition with the lowest AID score for KM (partitions 1 and 12 in Table 2, respectively) result from GDD-agreement. And since we expect GDD-agreement and RGF-distance to work better than their topologically simpler counterparts, these results might imply that filtered networks are preferred, at least from this perspective.

We also study the robustness of our approach to network filtering. That is, we check whether the same combinations of network construction and clustering parameters that result in 'statistically significant' partitions for unfiltered networks also result in 'statistically significant' partitions for filtered matrices (the last two columns of Table 1), and vice versa (the last two columns of Table 2). In this context, we find that HIE is less robust than KM. Specifically, for HIE, only 1 out of 8 'statistically significant' unfiltered partitions remains 'statistically significant' after filtering (Table 1), and only 3 out of 11 'statistically significant' filtered partitions remain statistically significant without filtering (Table 2).

On the other hand, for KM, these results are 6 out of 9 and 6 out of 10, respectively. Importantly, there is no partition that is both 'statistically significant' and semantically meaningful both with and without filtering.

Unfiltered vs. filtered networks: *summary.* Our results suggest that filtering of similarity matrices can significantly affect the topology of the resulting networks and the quality of the resulting partitions. Importantly, network construction and clustering strategies that work well for unfiltered networks do not necessarily work well for filtered network (and vice versa). Although filtering does not drastically improve the quality of the image partitions, it leads to a slightly larger number of 'statistically significant' and semantically meaningful partitions. Filtered networks also correlate more significantly than unfiltered networks with three out of four sets of 'ground truth' image distances. Furthermore, filtering works better in combination with more topologically constraining distance measures, as we would expect, which may imply that filtered networks have less noise compared with unfiltered networks.

3.4 *In-depth analysis and interpretation of a representative network-based image partition*

We now focus on a representative 'statistically significant' and semantically meaningful network-based partition for further in-depth analysis. In particular, we select partition 10 from Table 1. We choose this partition out of all partitions from Tables 1 and 2 as it mimics the closest the AV 'ground truth' partition in terms of the number of clusters (both partitions have nine clusters) as well as cluster sizes. As such, it enables a fair interpretation of its results with respect to the AV 'ground truth'.

To analyse this partition, we compute for each cluster its average valence score as well as its average arousal score over all images in the cluster, after assigning each image the score of 0, 0.5 or 1 for its low, medium or high arousal as well as for negative, neutral and positive valence, respectively. Figure 5 depicts these averages for each cluster on a two-dimensional arousal–valence space, while Fig. 6 visualizes the membership of images in each cluster. Ideally, we would hope for considerable similarities in average arousal–valence scores within each cluster and considerable differences in the scores across the clusters, as this would validate the correctness of our approach.

Indeed, this is what we observe (Figs 5 and 6). Clusters 2 and 7 contain images with negative valence, such as garbage, vomit, thrash in C2 and a cockroach, surgery, and a gun in C7. They also contain images with neutral valence, such as a mask and a dental exam in C2 and an electric outlet and men in C7. In general, mean valence scores for these clusters are highly similar (0.28 for C2 and 0.25 for C7), but these clusters can be distinguished via mean arousal values (0.33 for C2 and 0.5 for C7), with C7 containing images that are more emotionally arousing than C2. With one exception, neither cluster contains any positively valenced images.

Conversely, clusters C8 and C9 contain mostly positive-valenced with occasional neutral-valenced images, but they rarely contain negatively valenced images; mean valence scores are 0.64 and 0.67 for C8 and C8, respectively. Example images in these clusters include a couple, a watermelon, a tomato in C8 and a rabbit and surfers in C9. Once again, arousal discriminates these clusters, with images in C9 (e.g. surfers) being more arousing than images in C8 (e.g. tomato); mean arousal score in C9 of 0.50 is double that in C8 of 0.23.

Clusters C1 and C5 occupy similar positions in the arousal–valence space with mean arousal scores of 0.71 and 0.73 and mean valence scores of 0.62 and 0.59, respectively. Sample images in these clusters include fireworks, money, gold and clowns (which are arousing and have positive valence) along with

surgery, a roach on a slice of pizza and a starving child (which are arousing and have negative valence). Therefore, clusters C1 and C5 seem to capture physiological responses to arousal rather than valence.

Conversely, clusters C4 and C6 occupy similar positions in the arousal–valence space, but there are some marked differences among these clusters. Most notable is the fact that C4 contains an equal number of positively valenced (e.g. a butterfly or a baby) and negatively valenced images (e.g. a crying family or skulls), which leads to a mean neutral-valence profile (with score of 0.44). On the other hand, although C6 has a similar mean valence score of 0.5, it mainly contains neutral images (e.g. a bicyclist, cheerleaders or a pig), rather than a mix of positively- and negatively valenced images. Additionally, even though average arousal values for these clusters are also similar, C6 is slightly more arousing (with mean arousal score of 0.39) than C4 (with mean arousal score of 0.28).

Finally, while cluster C3 is consistent with average neutral valence (with average valence of 0.44, containing neutral images of, e.g. a building, an actor, a man or a skyscraper), this cluster is more interesting in terms of the variability in arousal that it captures, since it contains highly arousing (e.g. erotic female), medium arousing (e.g. actor) and low arousing (e.g. a man) images, resulting in average medium–high arousal (with score of 0.61).

A detailed analysis of this exemplary partition unveils some interesting insights pertaining to physiological responses to affective stimuli. An interesting observation is that the physiological responses seem to compress the arousal–valence space rather than being uniformly distributed across the space (Fig. 5). This is because the clusters rarely adhere to the expected AV 'ground truth' in that there is rarely a cluster with images that perfectly map onto the AV 'ground truth' image classes. For example, there is

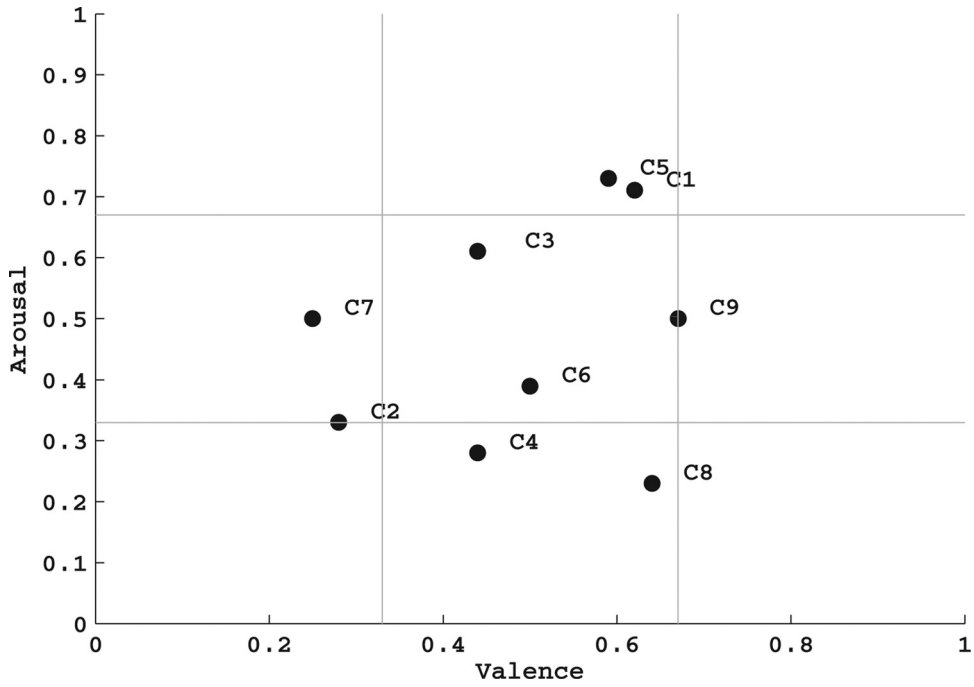

FIG. 5. Arousal–valence scores of different clusters in the representative network-based partition, where the scores are averaged over all images in the given cluster.

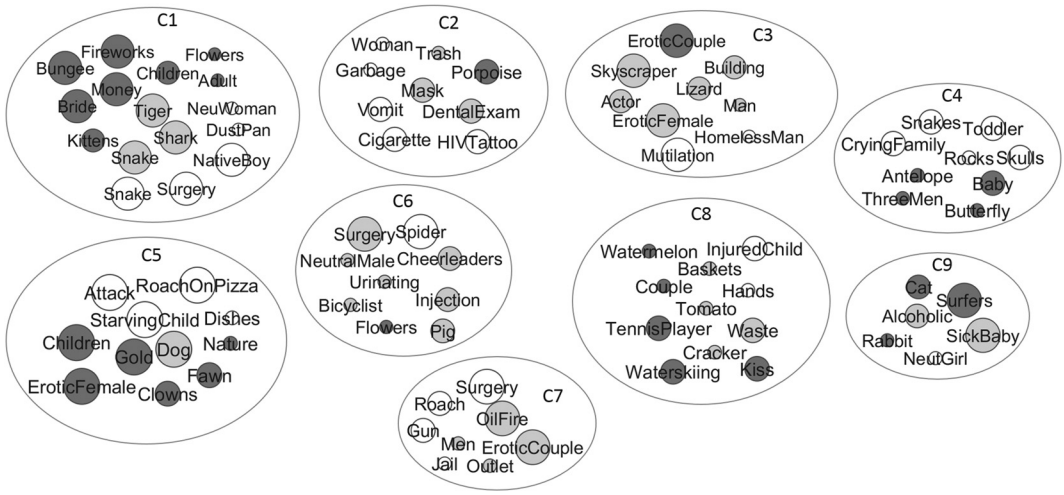

FIG. 6. Membership of the 89 images in different clusters of the representative network-based partition. Each image is depicted as a circle whose colour corresponds to its IAPS normative valence score (white—negative valence, light gray—neutral valence or dark gray—positive valence) and whose size corresponds to its IAPS normative arousal score (small—low arousal, medium—medium arousal or large—high arousal).

no high-arousal positive-valence cluster in our partition. Instead, the clusters contain blends of images that are mostly similar in valence but differ in arousal (e.g. C6), that are mostly similar in arousal but differ in valence (e.g. C5, C1), or that differ along both dimensions (e.g. C4). Given that our networks are created based on similarity of participants' physiological responses to images, our results suggest that inter-individual physiological responding does not neatly align with the expected AV 'ground truth', but is not entirely random either, as evident from the observed patterns in our clusters. We suspect that the 'meaning' of the images might also play a role in how individuals physiologically respond to them. Hence, both the affective dimension (i.e. valence and arousal) and the cognitive dimension (i.e. meaning) might be needed to explain variability in physiological responding. This is particularly encouraging since we have shown that our network-based approach captures both of these aspects, since it significantly aligns with both the AV 'ground truth' and LSA (meaning) partitions.

4. Conclusions

We use a network approach to study affective physiological data. Namely, we model images as networks and group images with similar network topologies. We perform a systematic analysis of the effect of different network construction and clustering approaches, concluding that the each choice can affect the results. For network inference and clustering communities, this highlights the importance of considering various strategies. Nonetheless, we show that via network analysis we can construct image partitions that are significantly similar to the AV 'ground truth' partition, while at the same time yield deeper insights by also being sensitive to semantics as estimated by LSA. 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 yield deeper insights by improving analysis of the data.

We introduce a framework for systematic network analysis of human physiological responses. There are several future extensions of our research. A next step could be to identify communities of individuals with respect to their (dis)similarities in physiological responses to affective stimuli and to investigate whether these communities can be discriminated on the basis of trait-based individual differences (e.g. personality factors, distress tolerance, generalized anxiety, etc.). Also, while our data are based on 18 subjects, studying more subjects would make the results more generalizable. While we construct image networks modelling similarities between responses of different individuals to a given image, one can also construct subject networks, modelling similarities between responses of a given subject to different images, which could give complementary insights. Finally, 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. In this context, the different signals could be studied individually from network perspective and then their results could be integrated for a more comprehensive understanding, or the networks corresponding to the different signals could be integrated first, prior to any network analysis; how exactly this should be done is the subject of future research, as it falls under the umbrella of the field of heterogeneous (network) data analysis, which is somewhat in its infancy. All of the above future extensions of our current work could lead to the ultimate goal of understanding how different emotions are manifested in physiological signals both within and across individuals.

Funding

This work was supported by the National Science Foundation [CCF-1319469 to T.M., EAGER CCF-1243295 to T.M., HCC-0834847 to S.D.M., DRL-1235958 to S.D.M.].

References

1. Hulovatyy, Y., D'Mello, S., Calvo, R. A. & Milenković, T. (2013) Network analysis improves interpretation of affective physiological data. *2013 International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pp. 470–477.
2. Watts, D. & Strogatz, S. (1998) The small world problem. *Collect. Dyn. Small-World Netw.*, **393**, 440–442.
3. Milenković, T. & Pržulj, N. (2008) Uncovering biological network function via graphlet degree signatures. *Cancer Inf.*, **6**, 257–273.
4. Newman, M. (2009) *Networks: An Introduction*. Oxford: Oxford University Press.
5. Meng, L., Milenković, T. & Striegel, A. (2014) Systematic dynamic and heterogeneous analysis of rich social network data. *5th Workshop on Complex Networks (CompleNet)*.
6. Mitrović, M., Paltoglou, G. & Tadić, B. (2010) Networks and emotion-driven user communities at popular blogs. *Eur. Phys. J. B-Condens. Matter Complex Syst.*, **77**, 597–609.
7. Izard, C. (2010) The many meanings/aspects of emotion: definitions, functions, activation, and regulation. *Emotion Rev.*, **2**, 363–370.
8. Picard, R. (1997) *Affective Computing*. Cambridge, MA: MIT Press.
9. Calvo, R. A. & D'Mello, S. K. (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.*, **1**, 18–37.
10. Zeng, Z., Pantic, M., Roisman, G. & Huang, T. (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**, 39–58.
11. Picard, R., Vyzas, E. & Healey, J. (2001) Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 1175–1191.

12. Chanel, G., Kronegg, J., Grandjean, D. & Pun, T. (2006) Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals, *Multimedia Content Representation, Classification and Security* (B. Gunsel, A. Jain, A. M. Teklap & B. Sankur eds). Lecture Notes in Computer Science. Heidelberg, Berlin: Springer, vol. 4105, pp. 530–537.

13. AlZoubi, O., D'Mello, S. & Calvo, R. (2012) Detecting naturalistic expressions of nonbasic affect using physiological signals. *Affect. Comput.*, **3**, 298–310.

14. Larsen, J., Berntson, G., Poehlmann, K., Ito, T. & Cacioppo, J. (2008) *The Psychophysiology of Emotion*. New York, NY: The Guilford Press, pp. 180–195.

15. Ekman, P. (1992) An argument for basic emotions. *Cogn. Emotion*, **6**, 169–200.

16. Herbelin, B., Benzaki, P., Riquier, F., Renault, O. & Thalmann, D. (2004) Using physiological measures for emotional assessment: a computer-aided tool for cognitive and behavioral therapy. *Int. J. Disabil. Hum. Dev.*, **4**, 269–277.

17. Liu, C., Agrawal, P., Sarkar, N. & Chen, S. (2009) Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *Int. J. Hum.-Comput. Interact.*, **25**, 506–529.

18. van der Zwaag, M., Janssen, J. & Westerink, J. (2012) Directing physiology and mood through music: validation of an affective music player. *IEEE Trans. Affect. Comput.*, **4**, 57–68.

19. Kim, K., Bang, S. & Kim, S. (2004) Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Eng. Comput.*, **42**, 419–427.

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

21. Ćosić, K., Popović, S., Horvat, M., Kukolja, D., Dropuljić, B., Kovač, B. & Jakovljević, M. (2013) Computer-aided psychotherapy based on multimodal elicitation, estimation and regulation of emotion. *Psychiatr. Danub.*, **25**, 340–346.

22. Ćosić, K., Popović, S., Kukolja, D., Horvat, M. & Dropuljić, B. (2010) Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders. *CyberPsychol. Behav. Soc. Netw.*, **13**, 73–78.

23. Russell, J. (2003) Core affect and the psychological construction of emotion. *Psychol. Rev.*, **110**, 145–172.

24. Monkaresi, H., Hussain, M. & Calvo, R. (2012) Classification of affects using head movement, skin color features and physiological signals. *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2664–2669. IEEE.

25. Dionísio, A., Menezes, R. & Mendes, D. A. (2004) Mutual information: a dependence measure for nonlinear time series. *Phys. A Stat. Mech. Appl.*, **344**, 326–329.

26. Vinh, N. X., Epps, J. & Bailey, J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, **9999**, 2837–2854.

27. Feizi, S., Marbach, D., Médard, M. & Kellis, M. (2013) Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.*, **31**, 726–733.

28. Fortunato, S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.

29. Solava, R. W., Michaels, R. P. & Milenković, T. (2012) Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, **28**, 480–486.

30. Salton, G. & McGill, M. J. (1986) *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

31. Kuchaiev, O., Stevanović, A., Hayes, W. & Pržulj, N. (2011) GraphCrunch 2: software tool for network modeling, alignment and clustering. *BMC Bioinform.*, **12**, 1–13.

32. Pržulj, N., Corneil, D. G. & Jurisica, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

33. Pržulj, N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, 177–183.

34. Milenković, T., Lai, J. & Pržulj, N. (2008) GraphCrunch: a tool for large network analyses. *BMC Bioinform.*, **9**, 1–11.
35. Vinh, N. X., Epps, J. & Bailey, J. (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary?. *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1073–1080. ACM.
36. Landauer, T. K. & Dumais, S. T. (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*, **104**, 211–240.
37. Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.*, **27**, 209–220.
38. Andreassi, J. L. (2013) *Psychophysiology: Human Behavior & Physiological Response*. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A. IAPS identifiers

Nine classes of the AV 'ground truth' partition, illustrated in Fig. 2(a), contain images with the following IAPS identifiers: (1) negative valence, low arousal: 2039, 2104, 2440, 2722, 5130, 7040, 9260, 9291, 9331, 9390; (2) negative valence, medium arousal: 1111, 1270, 2095, 2456, 3301, 6241, 9005, 9320, 9440, 9831; (3) negative valence, high arousal: 1050, 1205, 2730, 3000, 3212, 3213, 6520, 7380, 9075; (4) neutral valence, low arousal: 2190, 2397, 2499, 2720, 5875, 6150, 7041, 7255, 7287, 9700; (5) neutral valence, medium arousal: 1122, 1350, 2034, 2752, 2770, 2780, 7079, 9469, 9582, 9594; (6) neutral valence, high arousal: 1113, 1302, 1726, 1931, 3211, 3302, 4008, 4604, 7640, 9230; (7) positive valence, low arousal: 1605, 1610, 1620, 2000, 2370, 2501, 5200, 5760, 5811, 7325; (8) positive valence, medium arousal: 1463, 1540, 1630, 1920, 2071, 2092, 2345, 2352, 8205, 8350 and (9) positive valence, high arousal: 2208, 2347, 4180, 4693, 5480, 8179, 8206, 8496, 8500, 8502.