Pattern Recognition Letters
journal homepage: www.elsevier.com

The Mutual Information Between Graphs

Francisco Escolano[a,**], Edwin R. Hancock[b], Miguel A. Lozano[a], Manuel Curado[a]

[a]Department of Computer Science and AI, University of Alicante, 03690, Spain
[b]Department of Computer Science, University of York, 03690, United Kingdom

ABSTRACT

The estimation of mutual information between graphs has been an elusive problem until the formulation of graph matching in terms of manifold alignment. Then, graphs are mapped to multi-dimensional sets of points through structure preserving embeddings. Point-wise alignment algorithms can be exploited in this context to re-cast graph matching in terms of point matching. Methods based on bypass entropy estimation must be deployed to render the estimation of mutual information computationally tractable. In this paper the novel contribution is to show how manifold alignment can be combined with copula-based entropy estimators to efficiently estimate the mutual information between graphs. We compare the empirical copula with an Archimedean copula (the independent one) in terms of retrieval/recall after graph comparison. Our experiments show that mutual information built in both choices improves significantly state-of-the art divergences.

1. Introduction

1.1. Motivation

One of the key elements for building a pattern theory is the definition of a set of principled dissimilarity measures between the mathematical structures underpinning the theory. For instance, in vectorial pattern recognition, one of the fundamental degrees of freedom of an information theoretic algorithm (for clustering, matching, classification and learning) is the choice of a divergence. There are some possibilities including mutual information, Kullback-Leibler, Bregman divergences, and so on (see Escolano et al. (2009) for a review).

The mutual information $I(X;Y)$ between two variables X and Y is very interesting since it captures high-order statistical dependencies between the variables. However, when these variables are graphs we must address two issues. Firstly, we must express graphs X and Y as random variables, beyond the simplistic model of Erdös-Rényi model. In such model a *random graph* is built by assigning a probability to the edges. However this model does not fully characterize the probability that a given graph (with a variable number of vertices) is observed. Secondly, since $I(X;Y) = H(X) + H(Y) - H(X,Y)$ we must estimate the Shannon entropy $H(.)$ of a graph. There are several approaches for estimating graph entropy. The most efficient entropy estimators rely on functionals aiming to quantify the amount of information flowing through the graph. For instance, in Bai and Hancock (2013) the state vector of the steady state random walk on the graph defines a discrete probability function on the nodes. The Shannon entropy of such a probability function yields $H(.)$. On the other hand, quantum walks probing is used in Torsello et al. (2014) for providing mixed quantum states known as density matrices. Following Passerini and Severini (2009), the von Neumann entropy (or quantum entropy) maps discrete (graph) Laplacians to quantum states: scaling the graph Laplacian by the inverse of the volume of the graph we obtain a density matrix whose entropy can be computed using the spectrum of the discrete Laplacian. More recently Han et al. (2012) have approximated the von Neumann entropy by formulating it in terms of node degrees.

The above methods for estimating graph entropy operate on the graph itself, i.e., they consider the graph as a coder of node/vertex dependencies and describe entropy in terms of its capability for diffusing information. However, in this paper we consider that a graph is a special type of random variable with a bounded number of nodes and/or edges and we model structural distortion in terms of a novel coding (transforming graphs into low-dimensional manifolds). Then, it is possible to exploit the apparatus of bypass entropy estimators (Neemuchwala et al. (2005b), Leonenko et al. (2008)). In fact, bypass estimators do

[**] Corresponding author: Tel.: +34-965903897; fax: +34-965903902; *e-mail:* sco@dccia.ua.es (Francisco Escolano)

not rely on estimating probability density functions but on Euclidean distances between vectorial patterns. This means that the Parzen approximation of the probability density function is no longer needed since entropy can be estimated directly from the samples.

On the other hand, the development of graph embeddings which map vertices to multi-dimensional spaces *bypasses the rigid discrete representation of graphs*. After being embedded, the associated multi-dimensional subspace must retain the rich topological information of the original representation. Many embeddings have been proposed so far: ISOMAP (Tenenbaum et al. (2000)), Heat Kernels (Xiao et al. (2010)) , Diffusion Maps (Lafon and Lee (2006)), Laplacian Eigenmaps (Belkin and Niyogi (2003)), Commute Times (Qiu and Hancock (2007)), Centered Normalized Laplacian (Robles-Kelly and Hancock (2007)) among others. Most of the these latter structure preserving embeddings (i.e. distances in the embedding are correlated with structural properties) establish a formal link between topology (usually encoded in spectral terms) and some kind of metric or dissimilarity measure in the subspace. Understanding and exploiting the latter formal link is key to quantifying the effectiveness of the corresponding embedding for a given task. For instance, graph comparison. In Escolano et al. (2011) there are experimental graph comparisons showing that the Commute Time (CT) embedding outperforms the alternatives in terms of retrieval/recall for the best dissimilarity measure in a given set. In addition, the fact that the latter embedding induces a metric allows us to work in the multi-dimensional subspace of the embedding. Here, problems such as finding graph prototypes are more tractable. It is then possible to return to the original topological space via inverse embedding (Escolano and Hancock (2011)).

1.2. Contribution

With these ingredients at hand (bypass estimators and suitable embeddings), the mutual information between graphs can be defined in terms of *structural information channels* (section 2). In such channel model, there will be embedding-based encoders and inverse embedding decoders. The channel will be characterized by a conditional entropy relying on a global non-rigid transformation between the input embedding and the distorted one. We will devote Section 3 to present how to obtain a multi-dimensional estimation of Mutual Information (MI) from the combination of copulas and Rényi entropy estimators. In Section 4 we will compare MI for embedded graphs with other challenging dissimilarities. In order to perform a fair comparison we will use the GatorBait database which has been proven to be a very challenging one despite its small size. This is due to the fact that it exhibits we very high intra-class variability and very low inter-class variability in only 100 samples. In Section 5 we will present our conclusions and future work.

Our main contribution in this paper is to define graph similarity through a model of structural information channel where distortion relies on manifold and MI is estimated through different types of copula functions.

2. Information Channels and Manifold Deformation

Let $\mathcal{X} = (\mathcal{V}_X, \mathcal{E}_X)$ be a random variable $\mathcal{X} : \Omega \rightarrow E$ defined over the set of unweighted and undirected graphs Ω with node-sets $\mathcal{V}_X$ having $|\mathcal{V}_X| = n$ nodes. Then, its associated edge-set $\mathcal{E}_X \subseteq \mathcal{V}_X \times \mathcal{V}_X$ satisfies $|\mathcal{E}_X| \leq \binom{n}{2}$ and a realization of $\mathcal{X}$ is given by an $n \times n$ adjacency matrix $\mathcal{A}_X \in E$.
Let $K_X : \mathcal{V}_X \times \mathcal{V}_X \rightarrow \mathbb{R}$ be a topological similarity measure $K_X(i, j)$, ideally a kernel, between two nodes $i, j \in \mathcal{V}_X$. We assume that the probability mass $p(\mathcal{X})$ relies on the probability mass of $K_X(.,.)$ as follows: peaked similarity distributions yield less probable realizations than flat ones. This choice is convenient for two reasons. Firstly, it is consistent with recent definitions of graph entropy (see Passerini and Severini (2009), Escolano et al. (2012) and Han et al. (2012)). Secondly, it provides a principled framework for understanding graph distortion in terms of the distortions induced in $K_X(.,.)$.

Let C be an *structural information channel* $\mathcal{X} \rightarrow C \rightarrow \mathcal{Y}$ where $\mathcal{Y} = (\mathcal{V}_Y, \mathcal{E}_Y)$ satisfies $\mathcal{V}_Y = \mathcal{V}_X$. Then, the conditional probability $p(\mathcal{Y}|\mathcal{X})$ describes a noiseless channel with respect to the vertices or nodes, but a noisy channel with respect to the edges. The channel C generates structural noise (insertions and/or deletions of edges) through an unknown matching function $g : \mathcal{E}_X \rightarrow \mathcal{E}_Y \bigcup\{\Phi\}$, where Φ is the NULL label accordingly with Myers et al. (2000). Finding the function $g(.)$ is typically posed in terms of minimizing the graph-edit distance between $\mathcal{X}$ and $\mathcal{Y}$ (see Sanfeliu and Fu (1973)). Although many recent developments have proposed approximations of the graph-edit distance (see for instance Fischer et al. (2015)) they are (to some extent) rooted in marginalizing $p(\mathcal{Y}|\mathcal{X})$. Marginalization tends to capture or preserve local coherence between the matched edges at the cost of loosing global coherence, especially when the input graphs $\mathcal{X}$ and $\mathcal{Y}$ are unattributed.

Here, we propose a different approach which enforces global coherence. Let $f_X : \mathcal{V}_X \rightarrow \mathbb{R}^d$, with $d \ll n = |\mathcal{V}_X|$, a graph embedding function. The embedding $f_X(.)$ induces a manifold $\mathcal{M}_X$, i.e. a subspace of $\mathbb{R}^d$, where the structural similarities $K_X(i, j)$ between pairs of vertices $i, j \in \mathcal{V}_X$ are encoded by a geodesic. Graph embedding functions are such that the Euclidean norm $||f_X(i) - f_X(j)||^2$ is a reasonable approximation of the geodesic insofar d matches the intrinsic dimension of the manifold (see Escolano et al. (2011)).

Therefore, since a graph $\mathcal{X}$ is mapped to a subspace/manifold $\mathcal{M}_X \subseteq \mathbb{R}^d$ we assume that the embedding function $f_X(.)$ plays the role of an encoder associated with the channel C which transmits one manifold $\mathcal{M}_X$ at a time. Given a manifold to transmit, its encoding is not free of error, i.e. it is noisy: different vertices can be mapped to the same point of $\mathbb{R}^d$. However, we assume that the messages (resulting from the encoding) retain the global topology of their respective graphs $\mathcal{X}$. A simple model for the the conditional distribution $p(\mathcal{Y}|\mathcal{X})$ governing C is the usual factorization

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{i=1}^{n} p(\Theta_Y^{(i)}|\Theta_X^{(i)}) \; , \tag{1}$$

where $\Theta_Y^{(i)}$ and $\Theta_X^{(i)}$ are respectively the $i-$th points of manifolds

$\mathcal{M}_Y$ and $\mathcal{M}_X$. However, the above factorization is misleading, since we have

$$p(\Theta_Y^{(i)}|\Theta_X^{(i)}) \propto \exp\left\{-\frac{1}{2}\left\|\frac{\Theta_X^{(i)} - \mathcal{T}(\Theta_Y^{(i)};\mathbf{W})}{\sigma}\right\|^2\right\}, \qquad (2)$$

where $\mathcal{T}(.;\mathbf{W})$ is a *global* non-rigid transformation parameterized by $\mathbf{W}$, and σ is the bandwidth (see Escolano et al. (2011) for more details). This is consistent with assuming that we cannot observe the matching function $g(.)$ but instead its effects in the similarity matrix $K_X(.,.)$ in order to produce a new one $K_Y(.,.)$ which determines the embedding $f_{\mathcal{Y}} : \mathcal{V}_Y \rightarrow \mathbb{R}^d$ leading to $\mathcal{M}_Y$.

The framework developed in this paper encompasses our early research. A model for the information channel C does not only assume that an output manifold $\mathcal{M}_Y$ is received. It must also specify how it is decoded. We do that through an *inverse embedding*. In our previous work (see Escolano and Hancock (2011)) we showed that for certain types of embeddings, e.g. commute-time embeddings, it is possible to approximate $\mathcal{Y}$ with minimal error.

Consequently, in our model we naturally associate distortion (when the information rate exceeds the channel capacity) with excessive deformation, since the capacity of the channel, defined as $C = \max_{p(\mathcal{X})} I(\mathcal{X};\mathcal{Y})$, decays significantly with the increase of $\epsilon = \sum_{i=1}^n \left\|\Theta_X^{(i)} - \mathcal{T}(\Theta_Y^{(i)};\mathbf{W})\right\|^2$. This means that, although $\mathcal{T}(.;\mathbf{W})$ is chosen so that ϵ is minimized, the deformation is constrained by a regularization constant, i.e. the channel capacity is bounded by regularization.

Bridging deformation with mutual information $I(\mathcal{X};\mathcal{Y})$ opens up a way of analyzing structural pattern distortion in terms of rate-distortion theory. In the following section, we propose a means of estimating $I(\mathcal{X};\mathcal{Y})$ within this framework.

3. Mutual Information Between Graphs

3.1. Graphs as Random Variables

When heading $\mathcal{X} = (\mathcal{V}_X, \mathcal{E}_X)$ and $\mathcal{Y} = (\mathcal{V}_Y, \mathcal{E}_Y)$ as random variables, now with $|\mathcal{V}_X| = n, |\mathcal{V}_Y| = m$ with $m \neq n$ in general, we assume: (i) the existence of an upper bound B of the number of vertices for any graph encoded by a structural random variable, i.e. $n, m \leq B$; therefore the number of edges of any graph is bounded by $\binom{B}{2}$; (ii) the density mass $p(.)$ is defined according to a similarity measure $K(.,.)$ so that peaked similarity distributions yield less probable realizations than flat ones. These rules are followed by the graphs $\mathcal{X} = (\mathcal{V}_X, \mathcal{E}_X)$ feeding the structural information channel C.

In addition, the information channel C can also incorporate nodal noise as well (when $m \neq n$). The impact of this fact in the design of $p(\mathcal{Y}|\mathcal{X})$ is that we must establish a correspondence function (or matching field) $c : \mathcal{V}_Y \rightarrow \mathcal{V}_X$, where $c(.)$ is not necessarily a one-to-one matching. Then we can reformulate the conditional probability in terms of

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{u=1}^{m} p(\Theta_Y^{(u)}|\Theta_X^{c(u)}) , \qquad (3)$$

where the correspondence function $c(.)$ comes from the minimization of

$$\epsilon' = \sum_{i=1}^{n}\sum_{u=1}^{m} \left\|\Theta_X^{(i)} - \mathcal{T}(\Theta_Y^{(u)};\mathbf{W})\right\|^2 , \qquad (4)$$

for instance through regularized multi-dimensional point matching (see Myronenko and Song (2010)) and then compute the correspondence function from the optimal transformation $\mathcal{T}(.;\mathbf{W})$ minimizing ϵ'. After applying this transformation we have $c(u) = i$ for $u \in \mathcal{V}_Y$ and $i \in \mathcal{V}_X$. In this way, we ensure that the number of matched points is always $n = |\mathcal{V}_X|$ in order to be consistent with the information-theoretic alignment framework used for images (see Viola and III (1997) and Neemuchwala et al. (2005a)). The correspondence function plays then the role of providing a common reference system for comparing the two manifolds $\mathcal{M}_X$ and $\mathcal{M}_Y$ after the optimal alignment.

Regarding the impact of nodal noise in the structure of the similarity matrix $K_X(.,.)$ in order to produce another similarity matrix $K_Y(.,.)$, we interpret this noise in terms of rewiring the path structure of the original graph beyond a simple editing of the edges. New nodes can be added to $\mathcal{V}_X$ or old nodes of $\mathcal{V}_X$ can be deleted and this may imply the appearance or the deletion of edges. However, since $p(\mathcal{Y}|\mathcal{X})$ relies on the global transformation $\mathcal{T}(.;\mathbf{W})$ we assume that nodal noise will have a significant impact on the conditional probability $p(\mathcal{Y}|\mathcal{X})$ insofar the structure of the obtained manifold $\mathcal{M}_Y$ differs from $\mathcal{M}_X$ after the optimal alignment.

Consequently we will formulate the mutual information $I(\mathcal{X};\mathcal{Y})$ in terms of manifold distortion after the optimal alignment.

We summarize our approach in Algorithm 1 where there are explicit calls to compute non-rigid deformations (see details in Algorithm 2 introduced here for the sake of reproducibility[1]). In this regard, the choice of non-rigid deformations instead of using rigid or affine deformations is the explicit addition of a regularization term. In Algorithm 2 this term is taking into account when computing the Green's function. The non-rigid transformation used in graph comparison is grounded in two principles. Firstly, the geometry of the subspaces $\mathcal{M}_X$ and $\mathcal{M}_Y$ is typically non Euclidean. This means that a rigid or affine alignment will potentially lead to low frequency (poorly discriminative) results, unless a small number of samples/nodes justifies the use of such a strong regularizer (either rigid or affine). Secondly, given the graphs $\mathcal{X}$ and $\mathcal{Y}$ we embed their topological information (purely structural) in $\mathbb{R}^d$. Then, the quadratic assignment problem associated with graph matching is linearized in the embedding space. The classical rectangle rule of Graduated Assignment (Gold and Rangarajan (1996)) imposes the constraint that if two nodes match then their matches are also adjacent to enforce the smoothness of the matching field. This is exactly the role of regularization in $\mathbb{R}^d$. Regularization in the non-rigid approach is less constrained than in the rigid or affine cases.

[1] MATLAB code and data for reproducing all the experiments in this paper can be found in http://sites.google.com/site/scohomepage/ and will be soon submitted to IPOL.

Algorithm 1 StructuralMI

```
**Input:** Graphs $\mathcal{X}, \mathcal{Y}$ and embedding function $f$
**Output:** $I(\mathcal{X};\mathcal{Y})$
$\Theta_X = f_x(\mathcal{V}_X); \Theta_Y = f_y(\mathcal{V}_Y);$
$[\Theta_{Y\to X}, c]$ =NonRigid($\Theta_X$,$\Theta_Y$) ;
Joint $2d$ variable: $\Theta_{Z_{XY}} = [\Theta_X^T(c)\ \ \Theta_{Y\to X}^T]^T$;
Set $\alpha \approx 1, \alpha \neq 1$;
Compute $I_\alpha(\Theta_X, \Theta_{Y\to X})$ from Eq. 19:

    $\hat{I}_\alpha(\Theta_X, \Theta_{Y\to X}) = -\hat{I}_\alpha(\Theta_X(c)) - \hat{I}_\alpha(\Theta_{Y\to X}) + \hat{I}_\alpha(\Theta_{Z_{XY}})$ ;

Set $I(\Theta_X, \Theta_{Y\to X}) = \hat{I}_\alpha(\Theta_X, \Theta_{Y\to X})$;

$[\Theta_{X\to Y}, c']$ =NonRigid($\Theta_Y$,$\Theta_X$) ;
Joint $2d$ variable: $\Theta_{Z_{YX}} = [\Theta_Y^T(c')\ \ \Theta_{X\to Y}^T]^T$;
Compute $\hat{I}_\alpha(\Theta_{X\to Y}, \Theta_Y)$:

    $\hat{I}_\alpha(\Theta_{X\to Y}, \Theta_Y) = -\hat{I}_\alpha(\Theta_{X\to Y}) - \hat{I}_\alpha(\Theta_Y(c')) + \hat{I}_\alpha(\Theta_{Z_{YX}})$ ;

Set $I(\Theta_{X\to Y}, \Theta_Y) = \hat{I}_\alpha(\Theta_{X\to Y}, \Theta_Y)$;

**return** $I(\Theta_X, \Theta_{Y\to X}) + I(\Theta_{X\to Y}, \Theta_Y)$;
```

In addition, given that graphs $\mathcal{X}$ and $\mathcal{Y}$ are considered random variables the estimation of $I(\mathcal{X};\mathcal{Y})$ implies that we should carefully consider the $d-$dimensional representation of the structural patterns and more importantly to existing methods which are capable of capturing the statistical dependences between such representations. This naturally leads to the fundamental concept of *copula*, i.e. the amount of high-order statistical dependence between a collection of variables (see Nelsen (1999)) and the equivalence between the negative entropy of the copula and the mutual information between the variables. Therefore, we commence from the definition of mutual information and then that of the copula itself to develop a computational method for its estimation in Section 3.3.

3.2. Rényi Mutual Information

Given the conditional probability $p(\mathcal{Y}|\mathcal{X}) = p(\mathcal{Y},\mathcal{X})/p(\mathcal{X})$ we have that mutual information can be posed in terms of $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y})$. This formulation is more suitable for the practical use of *bypass* entropy estimators than the usual entropy formula $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X})$. This is due to the fact that $H(\mathcal{X}) = \int_{R^d} p(\mathcal{X}) \log p(\mathcal{X}) d\mathcal{X}$ and this requires the estimation of the density function $p(\mathcal{X})$. The curse of dimensionality precludes the use of *plug-in* estimators, such as Parzen's windows, for high d. On the other hand, bypass estimators, which only rely on the samples themselves, scale reasonably well with d (see Benavent et al. (2009) for a discussion).

As a result, bypass estimators are useful when structural variables are coded by sets of $d-$dimensional vectors. This is the case, since we have $\mathcal{X} \to \Theta_X$, through the embedding $f_x(.)$, and $\mathcal{Y} \to \Theta_Y$ through $f_y(.)$. In addition, we have applied the optimal global transformation $\mathcal{T}(.;\mathbf{W})$ and obtained a correspondence field $c : \mathcal{V}_Y \to \mathcal{V}_X$.

Algorithm 2 NonRigid (from Myronenko and Song (2010))

```
**Input:** Sets of $d-$dimensional points $X$ and $Y$
**Output:** Deformed $Y$, Correspondence function $c : Y \to X$
Initialize $\mathbf{W} = 0, \sigma \propto \sum_{i,u} \|X^{(i)} - Y^{(u)}\|^2, \beta > 0, \lambda > 0$
Build Green's function $\mathbf{G}$, where $\mathbf{G}_{ab} = e^{-\frac{1}{2\beta^2}\|Y^{(a)}-Y^{(b)}\|^2}$
**repeat**
    Update transformation $\mathcal{T} = Y + \mathbf{GW}$
    E-step: Compute $\mathbf{P}$

        $\mathbf{P}_{ui} = \frac{\mathbf{P}_{ui}}{\sum_k \mathbf{P}_{uk} + h(\sigma, d)}$ where $\mathbf{P}_{ui} = e^{-\frac{1}{2\sigma^2}\|X^{(i)}-\mathcal{T}(Y^{(u)})\|^2}$

    M-step: Solve $\mathbf{W}$

        $(\mathbf{G} + \lambda\sigma^2 diag(\mathbf{P1})^{-1})\mathbf{W} = diag(\mathbf{P1})^{-1}\mathbf{P}X - Y$

    Update $\sigma$
**until** *Convergence*
Obtain $c$: $c(u) = i$ if $\mathbf{P}_{ui} \approx 1$.

**return** $Y + \mathbf{GW}, c$;
```

Let Θ_X be the set of n $d-$dimensional points encoding $\mathcal{X}$ and $\Theta_{Y\to X}$ be their corresponding n points from Θ_Y after the global deformation and the correspondence field are applied. In Algorithm 1 we have used the notation $\Theta_X(c)$ to denote the samples of Θ_X corresponding with those of $\Theta_{Y\to X}$ via the correspondence mapping $c(.)$. However we drop this notation here for the sake of clarity.

Then, we have that

$$I(\Theta_X, \Theta_{Y\to X}) = H(\Theta_X) + H(\Theta_{Y\to X}) - H(\Theta_X, \Theta_{Y\to X}) , \quad (5)$$

is a proxy we use for $I(\mathcal{X};\mathcal{Y})$ once the global transformation $\mathcal{T}(.;\mathbf{W})$ and the correspondence field $c : \mathcal{V}_Y \to \mathcal{V}_X$ are found.

Obtaining $I(\Theta_X, \Theta_{Y\to X})$ with bypass entropy estimators for $d > 1$ is an open problem. The underpinning principle of most of the state of the art bypass estimators is that the Shannon entropies $H(\Theta_X)$, $H(\Theta_{Y\to X})$ and $H(\Theta_X, \Theta_{Y\to X})$ can be estimated from the distribution of inter point distances between the points of the $d-$dimensional sets Θ_X, $\Theta_{Y\to X}$ and those of the $2d-$dimensional joint $(\Theta_X, \Theta_{Y\to X})$ respectively. However, there is some controversy attending to the statistical consistency of the estimators. For instance, although the Leonenko et al. estimator (Leonenko et al. (2008)) has been successfully used in Escolano et al. (2011) to compute the SNESV measure, some authors have recently pointed out that there exist several formal flaws which do not ensure weak convergence. In Pál et al. (2010) it is suggested to relax the original problem and estimate instead the *generalized Mutual Information* (Rényi or $\alpha-$order, with $\alpha > 1$) whose limit is the Shannon MI when $\alpha \to 1$.

Then, given a $d-$dimensional random variable Θ, we have both the generalized entropy H_α and the generalized MI I_α defined respectively as

$$H_\alpha(\Theta) \quad = \quad \frac{1}{1-\alpha} \log \int_{R^d} q(\Theta)^\alpha d\Theta$$

$$I_\alpha(\Theta) \quad = \quad \frac{1}{\alpha-1}\log\int_{R^d}\frac{q(\Theta)^\alpha}{\left(\prod_{i=1}^d q_i(\Theta^{(i)})\right)^{\alpha-1}}d\Theta\ , \tag{6}$$

where $q:\mathbb{R}^d\rightarrow\mathbb{R}$ is the joint probability density function and $q_i:\mathbb{R}\rightarrow\mathbb{R}$ are the marginals.

Given a discrete sample Θ of n $d-$dimensional vectors $\Theta^{(1)},\Theta^{(2)},\ldots,\Theta^{(n)}$ (coming in this case from the manifold) we can obtain consistent estimators of both $H_\alpha(\Theta)$ and $I_\alpha(\Theta)$. In this regard, estimation relies on building graphs (one node per sample) whose edges rely on the Euclidean distances between the nodes ($d-$dimensional points). For instance, *entropic spanning graphs* (Hero et al. (2002)) are Minimum Spanning Trees (MSTs) computed from the $p-$th powers of the L_2 distances between the points. There is a mathematical relation between the sum of all the $p-$th powers of the Euclidean distances associated to the edges in the MST and the Rényi entropy. Nearest-neighbor (kNN) graphs are more robust to outliers than MSTs and they are used for instance in Pál et al. (2010) where a given set of nearest neighbors may be specified by a set of integers S and then used to define directed edges.

The kNN graph $\mathcal{G}$ generalizes the choice of the $k-$th nearest neighbor of each point (when $|S|=1$) as follows: $\mathcal{G}=(\mathcal{V},\mathcal{E})$ where $\mathcal{V}=\{\Theta^{(1)},\Theta^{(2)},\ldots,\Theta^{(n)}\}$ and $e_{ij}\in\mathcal{E}$ if $\Theta^{(j)}$ is the j-th nearest neighbor of $\Theta^{(i)}$ and $j\in S$. It has been proved that a consistent estimator of the Rényi entropy is

$$\hat{H}_\alpha(\Theta)=\frac{1}{1-\alpha}\log\frac{L_p(\Theta)}{\gamma n^{1-p/d}}\ , \tag{7}$$

where $L_p(\Theta)=\sum_{e_{ij}\in\mathcal{E}}||\Theta^{(i)}-\Theta^{(j)}||^p$, $p=d(1-\alpha)$ and γ is a constant that can be estimated by generating a large sample X in $[0,1]^d$ and then setting $\gamma=L_p(X)/n^{1-p/d}$.

3.3. Mutual Information and Copula Entropy

The formal link between the estimation of entropy and that of mutual information is the concept of a *copula* (Nelsen (1999)). Given a $d-$dimensional random variable we define $F(\Theta)=C(F_1(\Theta_1),F_2(\Theta_2),\ldots,F_d(\Theta_d))$ where $C(.)$ is a *copula function*, that is, a joint c.d.f. (cumulative distribution function) $C:[0,1]^d\rightarrow[0,1]$ which encodes the dependence between c.d.f. *uniform* marginals. Since a random vector $(F_1(\Theta^{(1)}),F_2(\Theta^{(2)}),\ldots,F_d(\Theta^{(d)}))$ has uniform marginals, we have that the joint probability $Prob[F_1(\Theta^{(1)}),F_2(\Theta^{(2)}),\ldots,F_d(\Theta^{(d)})]$ is a copula function.

A nice property of copula functions is that

$$I(\Theta)=-H(c(F(\Theta)))\ , \tag{8}$$

i.e. the *mutual information is equivalent to the negative entropy of the p.d.f. c(.) of the copula function* (Ma and Sun (2011)). This is consistent with the rescaling property of mutual information $I(\Theta)=I(h(\Theta))$ if $h(.)$ is a strictly increasing function (Pál et al. (2010)). Since each $F_i(.)$ satisfies this property, then we can formulate the Rényi mutual information in terms of :

$$I_\alpha(\Theta)=-H_\alpha(F_1(\Theta_1),F_2(\Theta_2),\ldots,F_d(\Theta_d))\ . \tag{9}$$

Given this formal link, we can bypass the estimation of $c(F(\Theta))$ by computing the so called *empirical copula*. If our choice of the copula function is the multi-dimensional c.d.f., then the empirical copula is given by taking the union or concatenating all $\hat{F}_j$ for $j=1,\ldots,d$, where $\hat{F}_j$ is an estimator of F_j:

$$\hat{F}_j(\Theta_i)=\frac{1}{n}|R_i|\ ,\ R_i=\left\{\Theta_j^{(k)}\le\Theta_j^{(i)}:1\le k\le n\right\}\ , \tag{10}$$

where $\Theta_j^{(i)}$ is the $j-$th component of the $i-$th sample, and similarly for $\Theta_j^{(k)}$. Then $\hat{F}_j(\Theta_i)$ is the average *rank* of Θ_i, the number of samples in the $j-$th dimension smaller or equal than Θ_i. Consequently, the empirical copula is given by $(W_1,W_2,\ldots,W_n)$ where $W_i=(\hat{F}_1(\Theta_1^{(i)}),\hat{F}_2(\Theta_2^{(i)}),\ldots,\hat{F}_d(\Theta_d^{(i)}))^T$. Given the empirical copula, a consistent estimator of the Rényi mutual information between the n samples is

$$\hat{I}_\alpha(\Theta)=-\hat{H}_\alpha(W_1,W_2,\ldots,W_n)\ . \tag{11}$$

For instance, in Fig. 1 we show both a) the samples of a 2D mixture of 3 Gaussians and b) the samples corresponding to the empirical copula and their connection through the kNN graph where $S=\{k\}$ and $k=4$. Although the graph has more than 3 connected components (due to the choice of k) the empirical copula reflects the community structure of the mixture (there are no links between the large communities). We have estimated $\hat{I}_\alpha=0.1734$ with $\alpha=1-p/d=0.9894$. In practice, the Shannon mutual information is estimated by choosing $\alpha\approx 1$.

3.4. Archimedean Copulas vs Empirical Copulas

Given the $\hat{F}_j(\Theta_i)\in[0,1]$ in Eq. 10 as estimators of the marginal c.d.f. of the $i-$th sample along the $j-$th dimension, it is possible to define alternatives to the empirical copula (which is $n\times d$-dimensional) in order to perform more efficient computations for the estimation of mutual information. A particularly interesting family is that of the *Archimedean* copulas (McNeil and Nešlehová (2009)). These copulas are specially designed for summarizing the dependence structure in multiple dimensions and collapse it into a single variable since copulas $C(.)$ are functions that satisfy $C:[0,1]^d\rightarrow[0,1]$. Archimedean copulas have been used in biometrics (Cao et al. (2012)) whereas empirical copulas have been used for *ICA* (Pál et al. (2010)).

In the Archimedean context, instead of estimating the usual $Prob[F_1(\Theta^{(1)}),F_2(\Theta^{(2)}),\ldots,F_d(\Theta^{(d)})]$ (or bypassing it in practice) it is preferred to define the copula as follows:

$$C(U_1,\ldots,U_d)=\psi\left(\psi^{-1}(U_1)+\ldots+\psi^{-1}(U_d)\right)\ , \tag{12}$$

where $U_j=F_j(\Theta_j)$ for $j=1,\ldots,d$ and $\psi(.)$ is an *Archimedean generator*. An Archimedean generator $\psi(.)$ is a non-increasing continuous function $\psi:[0,\infty)\rightarrow[0,1]$ which satisfies $\psi(0)=1$, $\lim_{x\rightarrow\infty}\psi(x)=0$ and is strictly decreasing in the interval $[0,\inf\{x:\psi(x)=0\}]$. The inverse $\psi^{-1}(.)$ is defined as follows $\psi:(0,1]\rightarrow[0,\infty)$ and by convention $\psi(\infty)=0$ and $\psi^{-1}(0)=\inf\{x:\psi(x)=0\}$. In addition, $\psi(.)$ only defines a copula if $\psi(.)$ is $d-$monotone. The function $\psi(.)$ is $d-$monotone, with $d\ge 2$, in a given interval (a,b) if is differentiable up to order $d-2$ and the derivatives $\psi^{(k)}(.)$ satisfy

$$(-1)^k\psi^{(k)}(x)\ge 0\ \text{ for }\ k=0,1,2,\ldots,d-2\ . \tag{13}$$

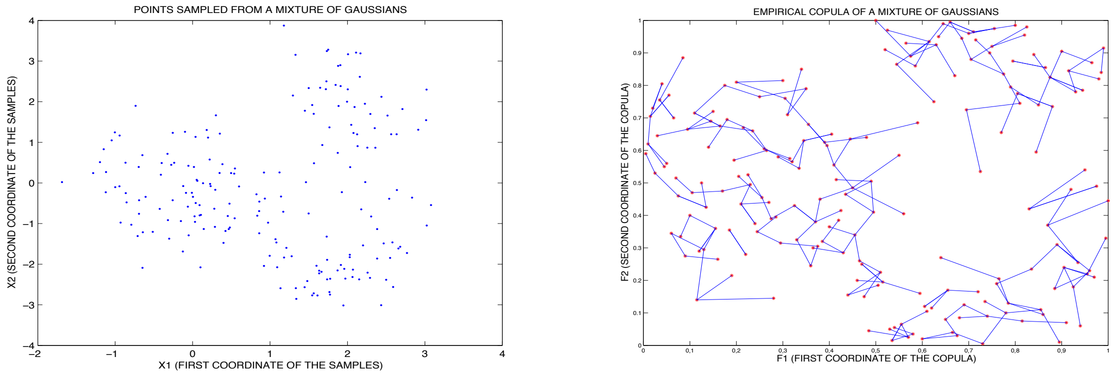

Fig. 1. Estimation of the empirical copula. Left: 200 **samples generated from a 2D Mixture of** 3 **Gaussians. Right: Empirical** $2D$ **copula and its associated** kNN **graph.**

for all $x \in (a,b)$ and also when $(-1)^{d-2}\psi^{(d-2)}(.)$ is both non decreasing and convex in (a,b). If the function has derivatives for all orders and the latter requirements are satisfied, then it is completely monotone. The notation $\psi_d(.)$ denotes a $d-$monotone function and $\psi_\infty(.)$ a completely monotone one.

The latter notation is very useful for parameterizing families of Archimedean copulas. For instance, the generator of the Clayton copula family is

$$\psi_\theta(x) = (1+\theta x)^{-1/\theta} \ . \tag{14}$$

For $\theta > 0$ the Clayton generator is completely monotone and we have that

$$\psi_0(x) = \lim_{\theta\to 0}(1+\theta x)^{-1/\theta} = \exp(-x) \tag{15}$$

is the so called *independence copula* in any dimension.

In this paper we will focus on $\psi(x) = \exp(-x)$ (and consequently on $\psi^{-1}(x) = -\log(x)$) because they are simple and parameter independent. Then, given the choice of the independent copula we have to estimate:

$$\begin{aligned} C(U_1,\ldots,U_d) &= \exp\left(-(-\log(U_1) - \ldots - \log(U_d))\right) \\ &= \exp\left(\sum_{j=1}^{d}\log(U_j)\right) \\ &= \exp\left(\log\left(\prod_{j=1}^{d}(U_j)\right)\right) = \prod_{j=1}^{d}(U_j) \ , \end{aligned} \tag{16}$$

where the copula is given by the factorization of the marginal c.d.f.s. This means that in a multi-dimensional setting the common structural information of the n samples is encoded independently by each of these marginals. For instance, setting $\hat{U}_j^{(i)} = \hat{F}_j(\Theta_i)$ implies that rank computation along the $j-$th dimension contains the structure of the samples w.r.t. Θ_i along that dimension. Therefore, let $V_i = C(\hat{U}_1^{(i)},\hat{U}_2^{(i)},\ldots,\hat{U}_j^{(i)}) \in [0,1]$ be new variables using now Eqs. 12 and/or 16 for defining $C(.)$. Then, we have that

$$\hat{I}_\alpha(\Theta) = -\hat{H}_\alpha(V_1,V_2,\ldots,V_n) \ , \tag{17}$$

that is, mutual information is estimated by a set of one-dimensional samples. Rényi estimation can be then done in time $O(n\log n)$. In fact, we can simplify the computations by avoiding the estimation of rank information. To do so, we take the original samples and normalize them so that they belong to $[0,1]^d$. In doing this, we are implicitly assuming that the value of each normalized sample component is a marginal c.d.f.. We then exploit Sklar's theorem: given a copula $C : [0,1]^d \rightarrow [0,1]$ and c.d.f. marginals $F_j(\Theta_j)$ then $C(F_1(\Theta_1),F_2(\Theta_2),\ldots,F_d(\Theta_d))$ defines a $d-$dimensional cumulative distribution function. This theorem allows us to use an Archimedean copula C' for defining $V_i' = C'(\mathcal{N}(\Theta_1^{(i)}),\mathcal{N}(\Theta_2^{(i)}),\ldots,\mathcal{N}(\Theta_j^{(i)}))$, where $\mathcal{N}(\Theta_j^{(i)})$ is the normalization of $\Theta_j^{(i)}$. As a result we have

$$\hat{I}_\alpha(\Theta) = -\hat{H}_\alpha(V_1',V_2',\ldots,V_n') \ , \tag{18}$$

and we can refer to this approach as the *raw estimation* of the Archimedean copula and, thus, the raw estimation of mutual information.

3.5. Mutual Information between Graphs

Given the above estimators of MI in terms of copulas, e.g. $\hat{I}_\alpha(\Theta) = -\hat{H}_\alpha(V_1',V_2',\ldots,V_n')$, we have a formal means of computing the proxy

$$I(\Theta_X,\Theta_{Y\to X}) = H(\Theta_X) + H(\Theta_{Y\to X}) - H(\Theta_X,\Theta_{Y\to X}) \ .$$

Since $\hat{I}_\alpha(\Theta_X) = -\hat{H}_\alpha(\Theta_X)$ and $\hat{I}_\alpha(\Theta_{Y\to X}) = -\hat{H}_\alpha(\Theta_{Y\to X})$, we only need to compute the estimation of the joint entropy $H(\Theta_X,\Theta_{Y\to X})$. In order to do this, we define the variable $\Theta_{Z_{XY}} = [\Theta_X^T \ \Theta_{Y\to X}^T]^T$ (the $2d$-dimensional concatenation of the two variables). The samples of $\Theta_{Z_{XY}}$ are obtained from the pairs defined by the correspondence field. Then, we have:

$$\hat{I}_\alpha(\Theta_{Z_{XY}}) = -\hat{H}_\alpha(\Theta_{Z_{XY}}) = -\hat{H}_\alpha(\Theta_X,\Theta_{Y\to X}) \ .$$

and also

$$\hat{I}_\alpha(\Theta_X,\Theta_{Y\to X}) = -\hat{I}_\alpha(\Theta_X) - \hat{I}_\alpha(\Theta_{Y\to X}) + \hat{I}_\alpha(\Theta_{Z_{XY}}) \ , \tag{19}$$

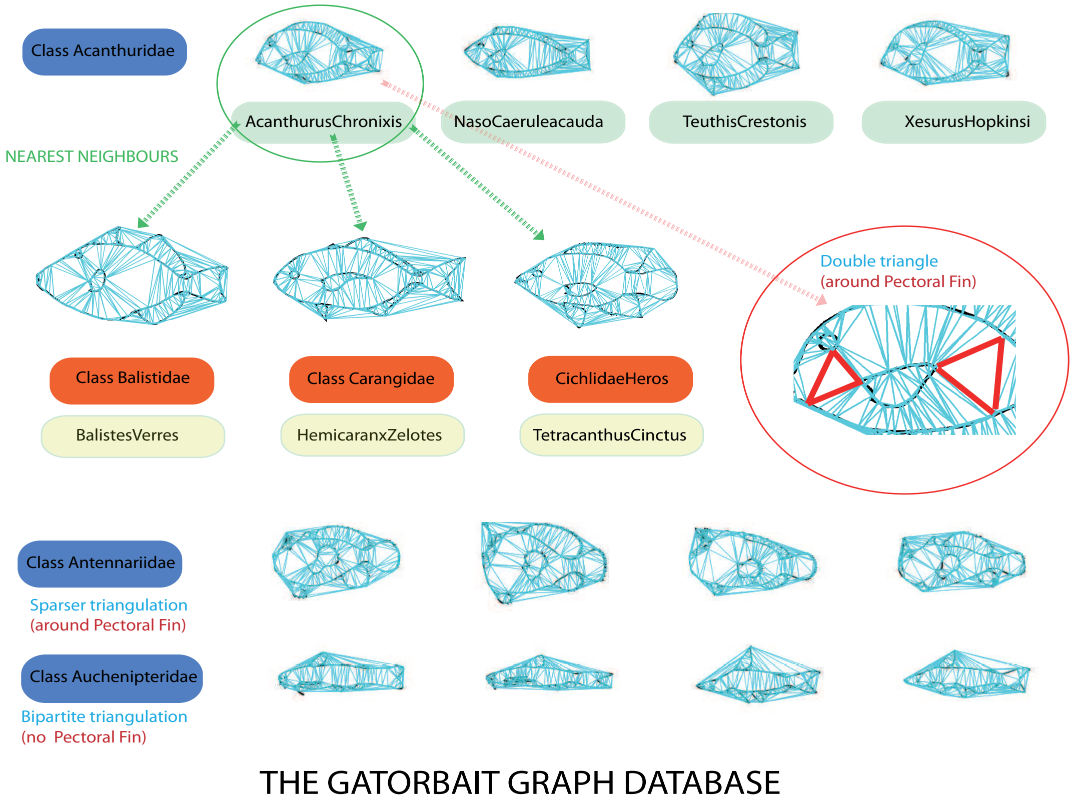

Fig. 2. Summary of the GatorBait Graph Database. The dataset has 100 Delaunay triangulations distributed in 30 classes. Classes are associated with fish genus and therefore we have high intra-class variability. For instance, taking as principal topological feature the distribution of triangules around the pectoral fin, we have that this feature varies among the graphs belonging to the same class. Different triangulating topologies characterize different classes. However, there is also a high inter-class variability since the *kNN* of a given graph belong frequently to different classes.

so that

$$I(\Theta_X, \Theta_{Y\to X})) = \lim_{\alpha\to 1} \hat{I}_\alpha(\Theta_X, \Theta_{Y\to X}) \,. \tag{20}$$

However, the above measure is not symmetric with respect to the non-rigid transformation applied to the data, and in general we have that $I(\Theta_X, \Theta_{Y\to X})) \neq I(\Theta_{X\to Y}, \Theta_Y))$. Therefore, the proxy of the *mutual information* between two graphs $\mathcal{X}$ and $\mathcal{Y}$ is given by the symmetrization:

$$\hat{I}_\alpha(\Theta_X; \Theta_Y) = \hat{I}_\alpha(\Theta_X, \Theta_{Y\to X}) + \hat{I}_\alpha(\Theta_{X\to Y}, \Theta_Y) \,, \tag{21}$$

and

$$I(\Theta_X; \Theta_Y) = \lim_{\alpha\to 1} \hat{I}_\alpha(\Theta_X; \Theta_Y) \,. \tag{22}$$

In the following section we will analyze the discriminability of mutual information in two contexts, manely a) comparison of copula functions and b) comparison with state-of-the-art divergences/algorithms.

4. Experiments

We use the *GatorBait_100* database in our experiments. GatorBait has 100 shapes representing fishes from 30 different classes (see Fig. 2 where we summarize the main features of this graph dataset). These shapes are discretized and then their Delaunay triangulations (included in the publicly accessible UA Graph Database[2]) are retained for testing graph comparison/matching algorithms. In order to compare our MI measure with the one proposed in Escolano et al. (2011) (SNESV) through entropic manifold alignment as well as with the classical quadratic function now computed through the PATH algorithm (Zaslavskiy et al. (2009)), we reproduce here the same experimental conditions. For instance, embedding functions $f_x(.)$ and $f_y(.)$ rely on the commute-time embedding with $d = 5$, and the settings of the CPD (Coherent Point Drift) point-matching algorithm (see Myronenko and Song (2010)) are the same. The embedding dimension d, is typically bounded by the results obtained by classical estimators of the intrinsic dimensions of $\mathcal{M}_X$ and $\mathcal{M}_X$ (see Costa and Hero (2004)). Final adjustment of d is typically done in the proximity of $d = 5$ since the bypass entropy estimator is relatively robust to the curse of dimensionality.

4.1. *Comparison of Copula Functions*

In Fig. 3-(top left) we show the average recall/retrieval curves for several copula functions, namely a) the *Archimedean cop-*

[2] http://www.rvg.ua.es/graphs/dataset01.html

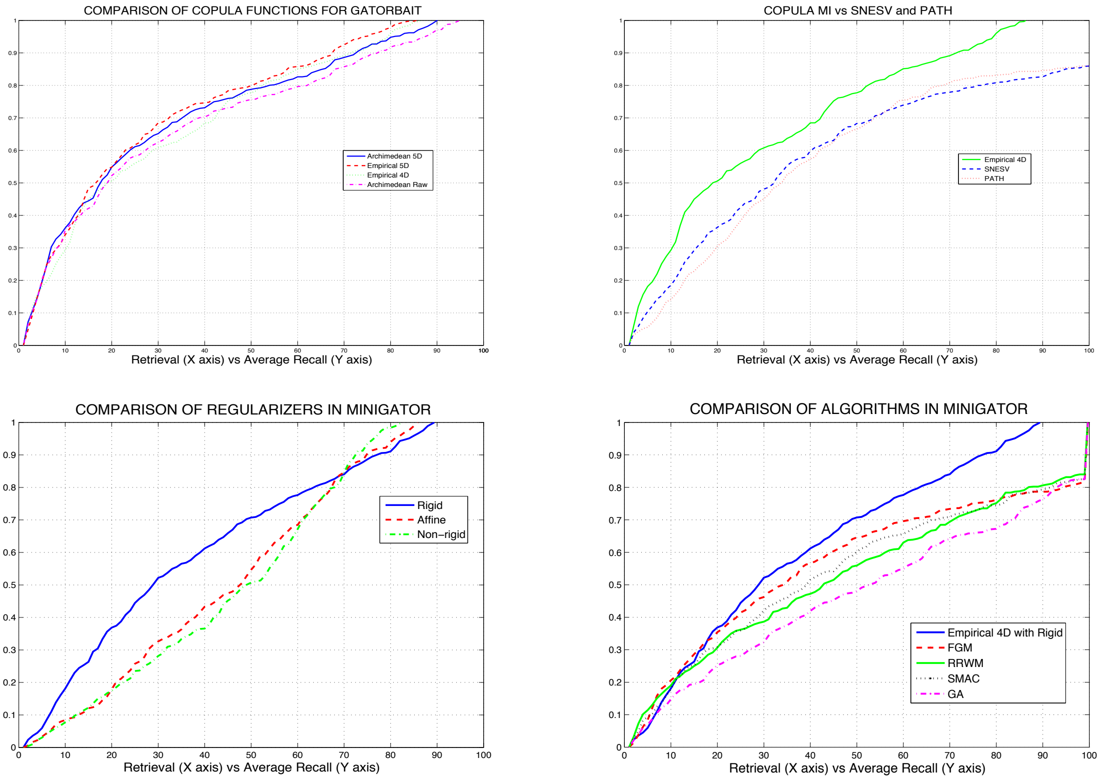

Fig. 3. Top: Retrieval-recall experiments in GatorBait.Top-left: comparison between different copula functions. Top-right: comparison with SNESV and PATH similarities. Bottom: Experiments in MiniGator. Bottom-left: Comparison between the rigid, affine and non-rigid regularizers (always using the 4D empirical copula. Bottom-right: Comparison between the rigid regularizer + 4D empirical copulas and similarities of state-of-the-art algorithms.

ula (independent) for $d = 5$, b) the *empirical copula* with $d = 5$, c) the *empirical copula* with $d = 4$, and finally d) the *raw Archimedean copula* with $d = 5$. The 5D empirical copula with AUC=75.1600 outperforms all the alternatives. It is followed by the 4D empirical copula with AUC=72.05. However, in this kind of recall/retrieval curve, for two similar AUCs the best curve is the one that grows faster. The 4D empirical copula intersects the alternatives (but the 5D empirical copula) when nearly 45 retrievals are performed. Therefore both Archimedean copula functions (the independent and the raw independent) with AUC=73.26 and AUC=70.60 outperform the 4D empirical copula. However, both Archimedean copula functions are outperformed by the empirical copula for the same dimensionality. This is consistent with the information fusion performed by the Archimedean functions which in fact produce faster (one-dimensional) Rényi estimators. In practice we may use the Arquimedean independent if we have time constrains in our structural recognition systems and this is better than reducing the dimensionality of the empirical copula.

4.2. *Comparison with SNESV and PATH*

In order to compare our approach with SNESV and the quadratic assignment function (without attributes) optimized by the PATH algorithm (both with $d = 5$) we will use the worst copula function, the 4D empirical copula. As we show in Fig. 3-(top right), even the worst copula significantly outperforms both SNESV and the quadratic assignment function optimized by PATH. This is consistent with the high-order statistical dependence information captured by mutual information. For example, in the case of SNESV, increasing the dimensionality leads to decrease the performance (for $d > 5$). However, when computing the mutual information we estimate a Rényi entropy for 2d =8 dimensions (the joint entropy). The joint entropy term is key for capturing the high-order dependencies and it is consistently estimated even for a high number of dimensions. On the other hand, SNESV (with AUC=59.60) outperforms PATH (with AUC=58.67). Actually PATH cuts SNESV at close to 55 retrievals. Entropic Alignment (SNESV) outperforms PATH and requires less intense computations than PATH.

4.3. *MiniGator and Comparison with Attributed Methods*

In order to study the robustness of our approach with respect to the number of samples (average size of the graphs) we have used a decimated version of *GatorBait_100* dubbed as *MiniGator*. To construct *MiniGator*, we retain 10% of the points of each shape in *GatorBait_100* and then build the associated Delaunay triangulation. To choose the most suitable regularizer (rigid, affine or non-rigid), our working hypothesis is that the smaller the number of samples the stronger the regularizer must be in order to minimize inter-class confusions. In Fig. 3-(bottom left) we show the performance curves for the three choices in MiniGator. In all cases $d = 5$ and the copula is the empirical one. We observe a significant performance degradation with respect to GatorBait. The best choice (rigid regularizer) provides an AUC of 65.20 (vs AUC=72.05 for the 4D empirical copula). In addition, the performance degrades significantly if we relax the regularizer (the affine regularizer does not come from a constrained optimization problem). Neither the affine nor the non-rigid choices are competitive.

Given the rigid regularizer (the best choice in MiniGator) we proceed to compare it with other state-of-the-art-algorithms, most of them relying on attributes, such as the Factorized Graph Matching(FGM), (see Zhou and la Torre (2012)), which actually is an attributed and optimized version of PATH. We also explore: Reweighted Random Walks Matching (RRWM), (see Cho et al. (2010)), Spectral Matching with Affine Constraint (SMAC), (see Cour et al. (2006)) and Graduated Assignment, (see Gold and Rangarajan (1996)). We show the performances obtained in Fig. 3-(bottom right). Our MI-based method, which is purely topological, outperforms all the alternatives and FGM is the second best choice.

5. Conclusions and Future Work

In this paper we have introduced a novel similarity measure for graph comparison through manifold (entropic) alignment: the mutual information (MI) between graphs. Estimating MI is addressed through the combination of copula functions and Rényi entropy estimators. We have studied both the empirical and the Archimedean copula functions. Empirical copula functions yield the best discriminative results, although the performance of Archimedean functions is very close to that of the empirical ones despite their formal structure. In addition, Archimedean copulas may be computed in sub-quadratic time, whereas empirical ones have a quadratic complexity. When compared with state-of-the-art similarities/algorithms the worst copula function outperforms very significantly the alternatives.

Future work includes the analysis of different families of copulas and the definition of ensembles of copulas. It also includes the formulation of a unified cost function for finding the alignment that maximizes mutual information. Since Rényi estimators rely on spanning trees or kNN graphs it is possible to approximate the derivatives of such estimators (it is straightforward for $k = 1$). In addition, the inclusion of novel and more efficient copulas opens new perspectives for the formalization of an information theory for graphs. For instance, we can study the implications of the channel coding theorem in graph theory and pattern recognition as well as having a better intuition of the meaning of entropy and coding in this context.

Acknowledgments

Funding. F. Escolano, M.A. Lozano: Project TIN2012-32839 (Spanish Gov.). M. Curado: BES-2013-064482 (Spanish Gov.). E. R. Hancock: Royal Society Wolfson Research Merit Award.

References

Bai, L., Hancock, E.R., 2013. Graph kernels from the jensen-shannon divergence. Journal of Mathematical Imaging and Vision 47, 60–69. URL: `http://dx.doi.org/10.1007/s10851-012-0383-6`, doi:`10.1007/s10851-012-0383-6`.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15, 1373–1396.

Benavent, A.P., Ruiz, F.E., Sáez, J.M., 2009. Learning gaussian mixture models with entropy-based criteria. IEEE Trans. Neural Networks 20, 1756–1771. URL: `http://dx.doi.org/10.1109/TNN.2009.2030190`, doi:`10.1109/TNN.2009.2030190`.

Cao, D., Chen, C., Adjeroh, D., Ross, A., 2012. Predicting gender and weight from human metrology using a copula model, in: Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on, pp. 162–169. doi:`10.1109/BTAS.2012.6374572`.

Cho, M., Lee, J., Lee, K., 2010. Reweighted random walks for graph matching, in: Proc. of ECCV.

Costa, J., Hero, A., 2004. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. IEEE Transactions on Signal Processing 52, 2210–2221.

Cour, T., Srinivasan, P., Shi, J., 2006. Balanced graph matching, in: Proc. of NIPS.

Escolano, F., Hancock, E., 2011. From points to nodes: Inverse graph embedding through a lagrangian formulation, in: CAIP (1), pp. 194–201.

Escolano, F., Hancock, E., Lozano, M., 2011. Graph matching through entropic manifold alignment, in: CVPR, pp. 2417–2424.

Escolano, F., Hancock, E.R., Lozano, M.A., 2012. Heat diffusion: Thermodynamic depth complexity of networks. Phys. Rev. E 85, 036206. URL: `http://link.aps.org/doi/10.1103/PhysRevE.85.036206`, doi:`10.1103/PhysRevE.85.036206`.

Escolano, F., Suau, P., Bonev, B., 2009. Information Theory in Computer Vision and Pattern Recognition. Springer, Computer Imaging, Vision, Pattern Recognition and Graphics, New York.

Fischer, A., Suen, C.Y., Frinken, V., Riesen, K., Bunke, H., 2015. Approximation of graph edit distance based on hausdorff matching. Pattern Recognition 48, 331–343. URL: `http://dx.doi.org/10.1016/j.patcog.2014.07.015`, doi:`10.1016/j.patcog.2014.07.015`.

Gold, S., Rangarajan, A., 1996. A graduated assignment algorithm for graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence 18, 377–388.

Han, L., Escolano, F., Hancock, E.R., Wilson, R.C., 2012. Graph characterizations from von neumann entropy. Pattern Recognition Letters 33, 1958–1967. URL: `http://dx.doi.org/10.1016/j.patrec.2012.03.016`, doi:`10.1016/j.patrec.2012.03.016`.

Hero, A., Ma, B., Michel, O., Gorman, J., 2002. Applications of spanning entropic graphs. IEEE Signal Processing Magazine 19, 85–95.

Lafon, S., Lee, A., 2006. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1393–1403.

Leonenko, N., Pronzato, L., Savani, V., 2008. A class of renyi information estimators for multidimensional densities. Annals of Statistics 36, 2153–2182.

Ma, J., Sun, Z., 2011. Mutual information is copula entropy. Tsinghua Science & Technology 16, 51–54.

McNeil, A., Nešlehová, J., 2009. Multivariate archimedean copulas, d-monotone functions and 1-norm symmetric distributions. Annals of Statistics 37, 3059–3097.

Myers, R., Wilson, R.C., Hancock, E.R., 2000. Bayesian graph edit distance. IEEE Trans. Pattern Anal. Mach. Intell. 22, 628–635. URL: `http://doi.ieeecomputersociety.org/10.1109/34.862201`, doi:`10.1109/34.862201`.

Myronenko, A., Song, X.B., 2010. Point-set registration: Coherent point drift. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 2262–2275.

Neemuchwala, H., Hero, A., Carson, P., 2005a. Image matching using alpha-entropy measures and entropic graphs. Signal Processing 85, 277–296.

Neemuchwala, H., Hero, A.O., Carson, P.L., 2005b. Image matching using alpha-entropy measures and entropic graphs. Signal Processing 85, 277–296.

Nelsen, R., 1999. An Introduction to Copulas. Lecture Notes in Statistics, New York.

Pál, D., Póczos, B., Szepesvári, C., 2010. Estimation of r?enyi entropy and mutual information based generalized nearest-neighbor graphs, in: Proc. of NIPS.

Passerini, F., Severini, S., 2009. Quantifying complexity in networks: The von neumann entropy. IJATS 1, 58–67. URL: `http://dx.doi.org/10.4018/jats.2009071005`, doi:`10.4018/jats.2009071005`.

Qiu, H., Hancock, E., 2007. Clustering and embedding using commute times. IEEE Trans. on PAMI 29, 1873–1890.

Robles-Kelly, A., Hancock, E., 2007. A riemannian approach to graph embedding. Pattern Recognition 40, 1042–1056.

Sanfeliu, A., Fu, K., 1973. A distance measure between attributed relational graphs for pattern recognition. IEEE Trans. Systems Man Cybernet. 13, 353–363.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Torsello, A., Gasparetto, A., Rossi, L., Bai, L., Hancock, E.R., 2014. Transitive state alignment for the quantum jensen-shannon kernel, in: Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings, pp. 22–31. URL: `http://dx.doi.org/10.1007/978-3-662-44415-3_3`, doi:`10.1007/978-3-662-44415-3_3`.

Viola, P., III, W.W., 1997. Alignment by maximization of mutual information. International Journal of Computer Vision 24, 137–154.

Xiao, B., Hancock, E.R., Wilson, R.C., 2010. Geometric characterization and clustering of graphs using heat kernel embeddings. Image Vision Comput. 28, 1003–1021.

Zaslavskiy, M., Bach, F., Vert, J.P., 2009. A path following algorithm for the graph matching problem. IEEE Trans. Pattern Anal. Mach. Intell. 31, 2227–2242.

Zhou, F., la Torre, F.D., 2012. Factorized graph matching, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, pp. 127–134.