net4Lap: Neural Laplacian Regularization for Ranking and Re-Ranking

M. Curado, F. Escolano, M.A. Lozano
Department of Computer Science and AI
University of Alicante,Alicante, 03690, Spain
Email: {mcurado,sco,malozano}@dccia.ua.es

E.R. Hancock
Department of Computer Science
University of York York, YO10 5DD, UK
Email: erh@york.ac.uk

***Abstract*—In this paper, we propose `net4Lap`, a novel architecture for Laplacian-based ranking. The two main ingredients of the approach are: a) pre-processing graphs with neural embeddings before performing Laplacian ranking, and b) introducing a global measure of centrality to modulate the diffusion process. We explicitly formulate ranking as an optimization problem where regularization is emphasized. This formulation is a theoretical tool to validate our approach. Finally, our experiments show that the proposed architecture significantly outperforms state-of-the-art rankers and it is also a proper tool for re-ranking.**

I. Introduction

The unsupervised learning of continuous/neural embeddings for words and objects [1], [2], [3], has attracted the interest of many researchers due to the experimental success of these representations. Since the introduction of `word2vec` (Skip-gram model) [1], there has been a growing interest in understanding its formal properties. This model infers local contexts by maximizing the correlation between the embeddings of neighboring words via SGD. The resulting embedding is encoded by the weights of the input layer of a shallow neural network (one hidden layer). Levy and Goldberg [2] looked at the co-ocurrence statistics and showed that the global optimum obtained with negative sampling is closely related to the factorization of the shifted PMI (Pointwise Mutual Information) matrix of the word-context probabilities. Pennington et al. [3] proposed the `GloVe` (Global Vector) model. `GloVe` is a weighted least-squares regression model, and the resulting embedding can be seen as a weighted MDS.

The relevance of word embeddings to this paper relies on their link with random walks [4] (generative models and topic transition). More precisely, Hashimoto et al. [5] formulated transition probabilities in terms of sub-Gaussian functions of Euclidean distances between words, thus linking word embeddings with manifold learning through Itô processes [6]. As we will formulate later on, the limiting log-transition probability converges to the geodesic in the manifold. More recently, Grover and Leskovec [7] have applied these ideas to propose `node2vec`, a method for inferring contextual feature embeddings from graphs and networks. The underlying mechanism is to simulate random walks to capture bags of paths that can feed the SGD, thus inferring geodesically consistent graph embeddings. This methodology has been tested for multi-label classification and link prediction in complex networks. However, herein we contribute with experimental evidence showing that `node2vec` cannot predict dense labellings such as the ones provided by Laplacian-based methods. Therefore, the power of neural embeddings is quite limited in problems involving regularization such as ranking on manifolds. However, another interesting result showed in this paper is that neural embeddings can boost the accuracy of Laplacian-based rankers.

Ranking is a well known problem. Given a graph (or affinity matrix) accounting for pairwise similarities between data on a manifold, and and query node, a ranker sorts relevant data to the query with respect to the global manifold structure [8], [9], [10]. Since ranking is closely related to semi-supervised labelling (transductive inference) [11], good rankers have been recenlty defined in terms of *minimizing the harmonic loss* [12], [13]. Harmonic losses quantify the lack of consistency between the ranking function and the structure of the manifold. Then, if the query node belongs to a given class, top ranked results must lie in the same class so that the harmonic loss is minimized. Ideally, all the elements of the class but the query must have a higher rank that the remainder nodes in the graph. To that end, ranking methods rely on diffusive (regularized) similarities that infer new links between nodes belonging to the same class.

However, the performance of ranking methods is heavily dependent on the quality of the input graph (e.g. KNN, Gaussian, ϵ-graphs). In this regard, many semi-supervised or supervised approaches have emerged along the last decade: RankBoost: [14] (combination of preferences), RankNet [15] (GD training by examples), minimization of ranking mistakes [16], Bipartite Ranking [17] (emphasis on positive and negative examples) and learning with SDP [18].

II. Contributions

In this paper, we propose a *neural-regularization ranking architecture*. We exploit both the flexibility and scalability of SGD to pre-process the input graph so that it is well conditioned for Laplacian regularization. As we show in Section III-B, neural embeddings tend to produce isotropic contexts. Although this representation is not able of doing effective ranking *per se*, its local isotropy is key for boosting the accuracy of rankers based on Laplacian regularization (see the experiments in Section IV). We also show that models of

random walks with return probabilities are proper samplers for bags-of-paths feeding the neural embedding (Section III-C). In addition, we formulate ranking as an optimization problem (Section III-D) as a means of formally validating our architecture. Finally, we introduce global centrality into a ranking approach. The role of global centrality is twofold: a) capture the underlying density of the manifold using a global measure, and b) increase the accuracy of re-ranking processes.

III. The net4Lap Model

A. The net4Lap Architecture

Given an input KNN graph, `net4Lap` (*neural networks for Laplacian-based regularization*), learns an embedding via SGD (see Fig. 1) from bags-of-paths sampled through random walks. A second KNN based on the embedding is more harmonic (locally isotropic) than the original and it feeds a Laplacian regularizer based on global centrality. As a result, a new KNN graph based on ranking relationships re-feeds the SGD neural model for re-ranking.

In the following, we describe the formal elements that implement the proposed architecture.

B. Local Isotropy of the Embedding

Given an undirected weighted KNN graph $G = (V, E, W)$, we have that $W_{ij} = h(\frac{1}{\sigma}||\mathbf{x}_i - \mathbf{x}_j||^2)$ are the pairwise affinities between the data (nodes) $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$, $h(.)$ is a sub-Gaussian function, and $(i, j) \in E$ if $W_{ij} > 0$. Then, SGD aims at inferring a function $f : V \to \mathbb{R}^d$ from:

$$\max_f \sum_{i \in V} \log \prod_{j:(i,j)\in E} Pr(j|f(i)) \tag{1}$$

where $Pr(j|f(i)) = e^{\langle f(i), f(j)\rangle)}/Z_i$ (log-probability proportional to correlation) and $Z_i = \sum_{k \in V} e^{\langle f(i), f(k)\rangle}$ is the local partition function [7]. Then, assuming that the $\mathbf{x}_i$ are clustered in classes $c \in \mathcal{C}$ and that similar data are mostly generated under similar discourses (classes) we have

$$Pr_{c\in\mathcal{C}}[(1-\epsilon)Z \leq Z_u \leq (1+\epsilon)Z] \geq 1 - \delta, \tag{2}$$

where $n = |V|$, $\epsilon = \tilde{O}(1/\sqrt{n})$, $\delta = \exp(-\Omega(\log^2 n))$, i.e. the partition function is *concentrated* [19]. This leads to $Pr(j|f(i)) \approx Pr(k|f(i))$ for common neighbors j, k of the node i. As a result, the entropy of the new weights $W'_{ij} = h(\frac{1}{\sigma'}||f(i) - f(j)||^2)$ is minimized wrt to that associated with the original W_{ij}s.

In this way, we obtain a new KNN graph $G' = (V, E', W')$ where the weights W'_{ij} are locally isotropic. Let then $g(\mathcal{S})$ be a ranking function applied to a subset $\mathcal{S} \subseteq V$ and $\mathcal{L}(\mathcal{S})$ be its *harmonic loss* defined as follows:

$$\mathcal{L}(\mathcal{S}) := \sum_{i\in\mathcal{S}, j\in\bar{\mathcal{S}}} W'_{ij}(g(i) - g(j)) \ . \tag{3}$$

Since the $W'_{ij}s$ are almost constant for $(i, j) \in E'$, minimizing $\mathcal{L}(\mathcal{S})$ leads to bound local variations of $g(.)$, i.e. $g(i) \approx \frac{1}{d'_i}\sum_{j:(i,j)\in E} g(i)W'_{ij}$, where d'_i is the degree of i. Harmonicity is thus enforced wrt the original KNN graph. This boosts significantly the accuracy of $g(.)$.

C. Role of Random Walks

Neural embeddings are build by sampling bags-of-paths in $G = (V, E, W)$, so that the context of any node i can be predicted from the statistical co-ocurrences with neighboring nodes. Sampling is driven by random walks (RWs). According to [5][6], if $P(X_t = j|X_{t-1} = i) = h\left(\frac{1}{\sigma}||\mathbf{x}_i - \mathbf{x}_j||^2\right)$ then

$$\lim_{t\to 0} -t \log P(X_t = \mathbf{x}_j|X_{t-1} = \mathbf{x}_i) \to \rho(\mathbf{x}_i, \mathbf{x}_j) \ , \tag{4}$$

where $\rho(.)$ is the geodesic. Since the above condition holds for classical RWs, where $p_{ij} = W_{ij}/d_i$, choosing them as path generators usually yields good empirical results on average (see Section IV). However, not all types of RWs work equally well. In particular, good alternative path samplers, such as Partial Absorbing RWs (PARWs) and `node2vec` RWs, exhibit an ability of slowing down the diffusion process.

PARWs [20] are defined as follows:

$$p_{ij} = \begin{cases} \frac{\alpha\lambda_i}{\alpha\lambda_i + d_i} & \text{if } i = j \\ (1 - p_{ii}) \times \frac{W_{ij}}{d_i} & \text{if } i \neq j \end{cases} \tag{5}$$

where the PARW gets absorbed in i with probability p_{ii}, $\lambda_i > 0$ modulates the mobility through the cluster (depending on its density) and $\alpha > 0$ plays an important role when using the PARW to define an affinity function (see next subsection).

On the other hand, `node2vec` relies on RWs with some *return* probability:

$$p_{ij} = \begin{cases} \frac{\pi_{ij}}{Z} & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\pi_{ij} = \alpha_{pq}(t, j) \times W_{ij}$, t is the last node visited by the RW and

$$\alpha_{pq}(t, j) = \begin{cases} \frac{1}{p} & \text{if } d_{tj} = 0 \\ 1 & \text{if } d_{tj} = 1 \\ \frac{1}{q} & \text{if } d_{tj} = 2 \end{cases} \tag{7}$$

where d_{tj} denotes the shortest path distance between nodes t and i, and p, q control, respectively how fast the RWs explores and leaves the neighborhood of a given starting node. More precisely, the walk tends to return to t either if p is large or it has many common neighbors. Setting q to a small value also helps to constrain the walk to a given neighborhood. Thus, the above RW is designed to explore a given graph in search of some structural properties such as homophily (inference of communities) and structural equivalence (nodes with the same role, such as hubs or between-cluster nodes). When applied to clustering, one must set a large p and/or a small q.

Finally, we have investigated MERWs [21] (Maximum Entropy RWs), recently used as a means of predicting visual saliency in computer vision [22]. Therefore, we emphasize the dissimilarities rather the affinities (otherwise, the RW travels mostly through intra-class links) and set $\hat{W}_{ij} = -\sigma \log W_{ij}$:

$$p_{ij} = \begin{cases} \frac{\hat{W}_{ij}}{\lambda} \times \frac{\phi_j}{\phi_i} & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where, λ is the Perron-Frobenius (dominant) eigenvalue of $\tilde{W}$ and ϕ is its associated eigenvector. This RW is designed

so that the entropy of the generative process increases at a rate $\log(\lambda)$, thus enforcing that all paths between different nodes are equally probable. This dependency of the global connectivity of the graph makes MERWs very appealing, however they tend to underestimate the geodesics, as we will show in the experiments.

D. Ranking as Laplacian-based Regularization

PARWs allow to define an interesting Laplacian-based similarity. This similarity relies on the probabilities $A = [a_{ij}]$ that PARWs starting at i get absorbed at j in finite time. In addition, $A = (\alpha\Lambda + L)^{-1}\alpha\Lambda$, where $L = D' - W'$ is the Laplacian of G' and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ (see [13],[20]).

However, A is not symmetric in general. It is row stochastic so that there are non-zero probabilities of being absorbed. Wu et al. [20] show that

$$\lim_{\alpha\to 0^+} (\alpha\Lambda + L)^{-1}\alpha\Lambda = \mathbf{1}\bar{\lambda}^T, \tag{9}$$

with $(\bar{\lambda})_i = \lambda_i / \left(\sum_{j=1}^n \lambda_j\right)$, regardless of graph structure. They use this fact to show that PARWs unify several models of RWs. Later, in [13], they focus on the left part of A, $M = (\alpha\Lambda + L)^{-1}$ to show that: a) it is a symmetric similarity matrix, and b) the choice of Λ determines how well the PARW moves through the manifold according to its local density. For instance, the harmonic loss predicts that $\Lambda = I$ is better for moving around dense clusters, whereas $\Lambda = D'$ is better when the manifold is locally sparse. A good empirical balanced choice is $\Lambda = H := diag(h_1, \ldots, h_n)$ where $h_i = \min(d'_i, median(\mathbf{d}'))$ and $\mathbf{d}' = diag(D')$ is the vector of degrees. In this regard, they denote Λ as a *regularizer* and use M as a ranking matrix, since it produces denser and stronger edges between nodes in the same class that those between nodes in different classes.

However, herein we make the regularizing power of M more explicit. More precisely, $M = \alpha\Lambda^{-1}A$, and A is the solution to the following minimization problem

$$\min_A Q(A) = ||\alpha\Lambda^{1/2}A - \Lambda^{1/2}||^2 + \gamma trace(A^T L A) \ . \tag{10}$$

with $\gamma = 1$. The right term penalizes large deviations associated with linked nodes:

$$trace(A^T L A) = \sum_{i,j} W'_{ij} \sum_k (a_{ik} - a_{jk})^2 \ . \tag{11}$$

For large W'_{ij} we have that the *differential absorption flow* $\sum_k (a_{ik} - a_{jk})^2$ induced by (i, j) must be kept as smaller as possible, thus forcing $a_{ik} \approx a_{jk}\ \forall k$. This differential flow can only grow when $W'_{ij} \approx 0$ or $(i, j) \notin E'$. This constraint is even harder in our ranking architecture, due to local isotropy (Section III-B). Therefore, in $G' = (V, E', W')$ (the KNN graph resulting from the neural embedding), the existence of an edge with a large W'_{ij} induces equally probable common absorption sites k for both i and j. As a result, the structure contained in $\{W'_{ij}\}$ imposes new affinities (links) based on the absorption probabilities.

The left term $T := ||\alpha\Lambda^{1/2}A - \Lambda^{1/2}||^2$, leads to minimizing a correlation. From the Frobenius norm, and setting $R = (A - I)$, we have that this term has the form

$$\begin{aligned} T &= trace\left(\left[\alpha\Lambda^{1/2}R\right]\left[\alpha\Lambda^{1/2}R\right]^T\right) \\ &= trace\left(\alpha^2\Lambda^{1/2}RR^T\Lambda^{1/2}\right) \\ &= trace\left(\alpha^2\Lambda RR^T\right) \\ &= \alpha^2 \sum_i \left[(a_{ii} - 1)^2 + \sum_k a_{ik}^2\right]\lambda_i \ . \end{aligned} \tag{12}$$

This leads to seek for self-absorption probabilities $a_{ii} \approx 1$ (it can be proved that $a_{ii} > a_{ij}$ for $j \neq i$). This is compatible with the minimization of the *absorption flow* $\sum_k a_{ik}^2$. The single stochasticity of A is implicit in the solution. In addition, if $\lambda_i = 1$ ($\Lambda = I$), the term T is less constrained that when $\lambda_i = d'_i$ ($\Lambda = D'$). This explains why PARWs surf very well through dense manifolds by setting $\Lambda = I$, whereas large absorptions are penalized when navigating through sparse manifolds ($\Lambda = D'$). Setting $\Lambda = H$ contributes with a clever trade-off that adapts ranking to the underlying manifold, as well as it relax our optimization problem.

E. Role of Global Centrality

Ranking based on PARWs is mostly focused on using *degree centrality*. Centrality characterizes the importance of a node within the graph, and degree is the most local measure of centrality. Wu et al. [13] expanded the M as an inverse, uncovering the following diffusive process:

$$M = \left(\sum_{k=0}^{\infty} [QW']^k\right) Q \ , \tag{13}$$

where degree D' plays a central role by normalizing the growing powers of W' with a descreasing weight $Q := (D' + \alpha\Lambda)^{-1}$. Let $q_i = d'_i + \alpha\lambda_i$. Then, each p-steps path $\Gamma = W'_{i_1,i_2}W'_{i_2,i_3}\ldots W'_{i_{p-1},i_p}$ is normalized by $\prod_{a=i_1}^{i_p} q_{i_l}$. If we set $\Lambda = D'$, this downweighting is harder than when we use $\Lambda = I$. This is important, even when setting $\alpha \to 0$.

Looking at the above expansion, we decided to explore the effect of building Λ in terms of a more global centrality. The shape of M suggested us to rely on one of the earliest global centrality measures, the one defined by Katz [23]. It is summarized as follows: *a node is important if it is linked to other important nodes*. In this way, an isolated high degree node is a false positive in terms of importance for the diffusion process implemented by M. Since isolated high degree nodes are usually associated to between-cluster nodes (inter-class noise) it is then key to implicity downweight the importance of these false positives.

Katz centrality is given by the vector $C = (I - \beta W')^{-1}\mathbf{1}$ where $\beta < 1/\lambda$ and λ is the main eigenvalue of W'. Then, similarly to the expansion of M, we have

$$(I - \beta W')^{-1} = I + \beta W' + \beta^2 W'^2 + \ldots = \sum_{k=0}^{\infty} \beta^k W'^k \ . \tag{14}$$

To commence, $(W'^k)_{ij}$ accounts for the weights of all paths of length k between nodes i and j. Then, all these entries are downweighted by a global quantity (a fraction of $1/\lambda^k$). From $C = (I - \beta W')^{-1}\mathbf{1}$, it is straight to obtain $C(i) = \sum_{k=0}^{\infty}\sum_{j=1}^{n}\beta^k(W'^k)_{ij}$. However, it is not so obvious that $C(i)$ relies on correlations between the i-th row of W' and the remainder rows (columns, since W' is symmetric).

Let $\mathbf{d}'_i$ denote the $i-$th row of W', d'_i the degree of node i and $\mathbf{d}' = diag(D')$ the vector of degrees. Then

$$C(i) = 1 + \beta d'_i + \beta^2 S_1 + \beta^3 S_2 + \dots \ , \qquad (15)$$

where $S_1 = [\mathbf{d}'_i\mathbf{d}'^T_1, \dots, \mathbf{d}'_i\mathbf{d}'^T_n]$ and $S_k = S_{k-1}\mathbf{d}$ for $k > 1$. For instance, in Fig. 2, we show the role of S_1. Since $\mathbf{d}'_i\mathbf{d}'^T_j$ retains the correlation between the *degree expansion* of nodes i and j, we have that central nodes are endowed with large correlations. For instance, $d_1 < d_j$ for all $j \in \{2, 3, 4\}$, but it is the most central node since it is linked to important nodes $2, 3$ and 4. In terms of correlations, we have $\sum_j \mathbf{d}'_1\mathbf{d}'^T_j > \sum_k \mathbf{d}'_i\mathbf{d}'^T_k \ \forall i \neq 1$.

Katz centrality is also an adaptive way of accounting for local manifold density. Back to Fig. 2, if $(2, 4) \in E$, then node 2 becomes the most central node: both its degree and its correlations grow. This is the typical scenario of dense manifolds (where $\Lambda = I$ is optimal). For sparse manifolds, where $\Lambda = D'$ is optimal, degrees decrease but correlations do not necessarily decrease, unless the W'_{ij}s are locally isotropic, as it is the case. As we show in Section IV, Katz centrality slightly improves the accuracy of ranking with respect to the adaptive use of I or D' when the inter-class noise is not too large. Setting $\lambda_i = C(i)$ and considering up to S_1 in Eq. 15, we have that Q in the expansion of M (Eq. 13) becomes

$$\begin{aligned} q_i &= d'_i + \alpha(1 + \beta d'_i + \beta^2 S_1) \\ &= \alpha + (1 + \alpha\beta)d'_i + \alpha\beta^2 \sum_j (\mathbf{d}'_i\mathbf{d}'^T_j)d'_j \ . \end{aligned} \qquad (16)$$

In a sparse manifold, $Pr[(i, j) \in E]$ is small for most of the nodes $j \neq i$. This leads to $q_i \approx \alpha + (1 + \alpha\beta)d'_i$ and this is compatible with the setting $\Lambda = D'$. However, in a dense manifold, the correlations $\mathbf{d}'_i\mathbf{d}'^T_j \approx n$ (become nearly constant and maximal) and therefore

$$\begin{aligned} q_i &\approx \alpha + (1 + \alpha\beta)d'_i + \alpha\beta^2 \times n \sum_j d'_j \\ &= \alpha + (1 + \alpha\beta)d'_i + \alpha\beta^2 n \times vol(G') \ , \end{aligned} \qquad (17)$$

where $vol(G')$ is the volume of the graph. The leading eigenvalue λ of W' satisfies $\max\{\bar{d}', \sqrt{d'_{max}}\} \leq \lambda \leq d'_{max}$. where $\bar{d}'$ is the average degree and d'_{max} is the maximum degree. In a dense manifold (and mostly under local isotropy) we have that $d'_i \approx \bar{d}'$ and $\lambda \approx d'_{max} \approx n$. Then

$$q_i \approx \alpha + (1 + \alpha\frac{1}{zn})d'_i + \alpha\frac{1}{zn} \times vol(G') \ , \qquad (18)$$

where $z \geq 1$ is the fraction of λ defining $\beta = 1/(z\lambda)$. As a result q_i is dominated by $vol(G')$. Therefore, setting $\Lambda = C$ (centrality) results in adaptive PARWs regarding the local density of the manifold. The trade-off given by setting $\Lambda = H_C$, with $H_C := diag(h_1, \dots, h_n)$ where $h_i = \min(C(i), median(C(i)))$, contributes to enforcing this adaptiveness.

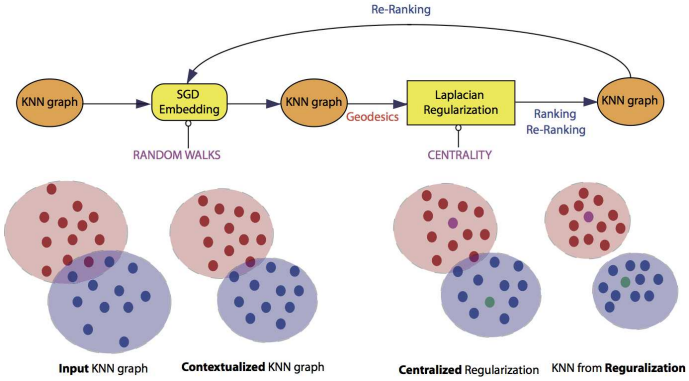

Fig. 1. `net4Lap`. Given a KNN graph, neural embedding (SGD with negative sampling) yields an harmonic version that feeds the Laplacian regularizer. As output, we obtain a denser graph suitable either for ranking or for obtaining an improved KNN graph which in turns feeds SGD for re-ranking.

Finally, we revisit the optimization problem formulated in Subsection III-D (Eq. 10) to explain why Katz centrality boots the re-ranking accuracy when feeding SDG with the modified M. Given the right term T (Eq. 12), we must minimize

$$\alpha^2 \sum_i \left[(a_{ii} - 1)^2 + \sum_k a_{ik}^2\right] C(i) \ . \qquad (19)$$

When local density is small, this term is as constrained as when setting $\Lambda = D'$. The main difference arises, however, when local density is large. Then $C(i)$ heavily depends on $vol(G') \gg d'_i$. This constraints the absorption probabilities much more in comparison with setting $\lambda_i = 1$. First, self-absorptions a_{ii} are amplified. Second, the absorption flow $\sum_k a_{ik}^2$ is minimized. Third, Eq. 11 shows that the differential absorption flow $\sum_k (a_{ik}^2 - a_{jk}^2)$ must be minimized as well in the neighborhood of an edge $(i, j) \in E$. Putting both absorption and differential absorptions together, we have that $a_{ik} \approx (1 - a_{ii})/n$, and similarly for a_{jk}. As a result, centrality-based ranking creates edges of very similar weights in the neighborhood of existing ones. In addition, one may think that when $W'_{ij} \approx 0$ (typically intra-class edges) then absorptions can be arbitrary large, but this is not possible due to the hard constraint imposed on the absorption flows. Therefore, centrality increases intra-class density whereas it reduces inter-class density. This behavior leads to pick mostly intra-class neighbors in the new KNN that has to feed the SGD during re-ranking. Boosting in terms of accuracy is due to the fact that the input to SGD is yet more locally isotropic that the original KNN graph $G = (V, E, W)$.

IV. Experiments

A. Datasets and Parameters

Our approach is tested in 4 datasets, each one with a particular distribution of inter-class noise. They are: NIST, Flickr32,

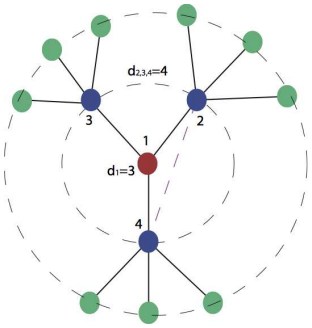

Fig. 2. Katz centrality. Node 1 is more centered than its neighbors despite it has a smaller degree than them: degree vectors correlations yield $\sum_j \mathbf{d}_1\mathbf{d}_j^T > \sum_k \mathbf{d}_i\mathbf{d}_k^T \ \forall i \neq 1$. However if the link $(2,4)$ exists, this is not true and node 1 becomes less central.

COIL-20 and CIFAR-10. In all cases, the sub-Gaussian is the neg-exponential: $h\left(\frac{1}{\sigma}||\mathbf{x}_i - \mathbf{x}_j||^2\right) = \exp\left(-\frac{1}{\sigma}||\mathbf{x}_i - \mathbf{x}_j||^2\right)$. Then, σ is obtained from fixing k (number of neighbors for the KNN) so that $\sigma = (k/100) \times S_{mean}$, where S_{mean} is the average Euclidean distance between pairs of nodes in the dataset.

Regarding the parameters of the neural network, when sampling of bags-of-paths with RWs, all of them have a fixed length $l = 80$ (there is not too much variation in accuracy if we increase this length). Another important parameter is the number of neurons in the hidden layer (dimesion d of the manifold). We found that $d = 128$ is a good choice. Reducing d usually reduces the accuracy. Finally, we set the window size for the context learner as in `node2vec`: $r = 10$.

NIST [24] is a subset of $n = 5,000$ examples, with $|C| = 10$ classes (500×10) of MNIST[1]: handwritten digits from 0 to 9 in images of 28×28 pixels. Their original dimensionality is reduced to $D = 86$ via PCA. Other parameters for ranking are: $\alpha = 0.01$, $k = 20$ for the KNN ($\sigma = 19.0371$).

This dataset exhibits a low intra-class noise and quite dense classes. The class corresponding to digit-1 has the largest affinities and the largest inter-class noise as well. Some other classes can be confused, such as digit-5, digit-7 and digit-9.

FlickrLogo32 [25] consists of 32 classes with 70 elements per class ($n = 2,240$): images of logos of different products[2]. Each image is represented by a GIST feature vector [26] ($D = 512$). Other ranking parameters: $\alpha = 0.05$, $k = 25$ for the KNN, ($\sigma = 0.2973$). The classes of this dataset are even denser than those of NIST but there is much more intra-class noise (both structured and unstructured).

COIL-20 The COIL-20 dataset[3] consists of 20 classes with 72 elements per class [27] ($n = 1,440$): images of objects taken from different points of view; their size is of 128×128 pixels ($D = 16,384$) . Other ranking parameters: $\alpha = 0.05$, $k = 20$ for the KNN graph ($\sigma = 2.08e + 07$). Half of the classes are very compact and free of inter-class noise. The remaining classes are less dense and exhibit a highly structured inter-class noise.

CIFAR is a subset of the CIFAR-10 dataset[4] with 10 classes and 100 elements per class ($n = 1,000$): images of 32x32 pixels (different from those used in [13]) represented by a GIST feature vector. Other parameters for ranking: $\alpha = 0.05$, $k = 25$ for the KNN ($\sigma = 0.2877$). All the classes are very sparse, even when representing images with the GIST descriptor, and inter-class links are more prominent than intra-class edges.

B. Evaluating the RWs

Our first experiment consists of evaluating the performance of different models of RWs. We want to evaluate to what extent our ranking approach is dependent on the choice of a particular model. In Table I we show two rows per dataset and type of RW (columns): in the first row we show the mean average precission (MAP) when degree centrality is used. In the second one we show the MAP when Katz centrality is applied (slightly better results with Katz in almost all cases). As we can see, RWs, PARWs and `node2vec` (for which we show the best p, q pairs) are proper choices, whereas MERWs are not adequate, due to their non-return and maximum entropy behaviors: equal probability of linking two nodes leads to be unable to discriminate between intra-class and inter-class links. Geodesics are then under-estimated. As a conclusion, either RWs or PARWs is a good choice, since `node2vec` requires learning the optimal p and q.

C. Ranking and Re-ranking Accuracies

In a second experiment (see Table II), we compare the MAPs for: a) Ranking with the neural embedding (column E), b) State-of-the-art ranking [13] *without neural embedding*, (H), c) ranking by applying H-ranking to the embedding (EH), d) same with Katz centrality (EK), e) Re-ranking based on EH-ranking (Re-EH) and f) same for EK-ranking (Re-EK).

The results validate our approach: 1) Neural embedding alone is not enough to achieve state-of-the-art MAPs, 2) State-of-the-art H-ranking is significantly improved (but in CIFAR-10) by feeding the regularization with the embedding; 3) Katz centrality slightly improves node centrality in ranking (EK vs EH), 4) However, in Re-ranking, Katz centrality clearly outperforms node centrality.

Our best ranking and re-ranking results are obtained with the NIST dataset (14% of gain in Re-ranking wrt H-ranking), since it has a small amount of inter-class noise. For Flickr the gain is reduced to 4%, and a similar gain is obtained for COIL. Finally, CIFAR is a very difficult dataset (sparse and inter-class noise), where H-ranking is slightly dominant.

Then, the predictions of the theory (local isotropy, constraining absorption probabilities through centrality, and relative invariance to the choice of random walks as samples, provided that they implement return probabilities) are validated by the experiments.

[1] http://yann.lecun.com/exdb/mnist/

[2] http://www.multimedia-computing.de/flickrlogos/

[3] http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[4] http://www.cs.toronto.edu/ kriz/cifar.html

TABLE I
EVALUATION OF RWS

	RW	PARW	$node2vec^{p,q}$	MERWa	MERWb
NIST	.7416	.7318	$\mathbf{.7438^{p=1,q=1}}$	.7236	.6338
	.7507	.7405	$\mathbf{.7529^{p=1,q=1}}$	.7332	.6383
Flickr	.5744	.5807	$\mathbf{.5817^{p=4,q=1}}$	.4137	.3855
	.5822	.5751	$\mathbf{.5832^{p=4,q=1}}$	.4122	.3858
COIL	**.7725**	.7669	$.7700^{p=1,q=1}$	.5879	.5171
	.7779	.7756	$.7656^{p=1,q=1}$	.5835	.5198
CIFAR	.2166	**.2173**	$.2171^{p=1,q=1}$	.2112	.2031
	.2180	**.2189**	$.2182^{p=1,q=1}$	.2130	.2026

TABLE II
MAP: RANKING AND RE-RANKING

	E	H	EH	EK	Re-EH	Re-EK
NIST	.5629	.6415	.7438	.7529	.7577	.7779
Flickr	.4978	.5433	.5817	.5832	.5792	.5839
COIL	.6960	.7336	.7725	.7779	.7588	.7787
CIFAR	.1511	.2242	.2173	.2189	.2131	.2140

V. CONCLUSION

In this paper, we have introduced `net4Lap`, a novel architecture for Laplacian-based ranking. The novelty of this architecture relies on: a) including shallow neural networks in the loop, b) implementing a theoretical framework for explaining why the neural-Laplacian combination boosts the performance of state-of-the-art rankers (induction of local isotropy and linking it with the harmonic loss), c) determining the nature of the random walks used for sampling bags-of-paths of fixed length, d) formulation of ranking in terms of a minimization problem where Laplacian regularization plays a fundamental role for theoretically validating the proposed architecture, d) injecting a global centrality measure in the ranking process, which is both consistent with the theory and plays a critical role in re-ranking. Finally, our experiments show that `net4Lap` outperforms the state-of-the-art both in ranking and re-ranking.

ACKWNOLEDGEMENTS

M. Curado, F. Escolano and M.A. Lozano are funded by the project TIN2015-69077-P of the Spanish Government.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.

[2] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2177–2185.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2014, pp. 1532–1543.

[4] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings," *CoRR*, vol. abs/1502.03520, 2015.

[5] T. B. Hashimoto, D. Alvarez-Melis, and T. S. Jaakkola, "Word embeddings as metric recovery in semantic spaces," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 273–286, 2016.

[6] T. B. Hashimoto, Y. Sun, and T. S. Jaakkola, "From random walks to distances on unweighted graphs," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2015, pp. 3429–3437.

[7] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.

[8] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *Advances in Neural Information Processing Systems 16, NIPS*, 2003, pp. 169–176.

[9] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, 2007.

[10] X. Zhou, M. Belkin, and N. Srebro, "An iterated graph laplacian approach for ranking on manifolds," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 877–885.

[11] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, ser. NIPS'03, 2003, pp. 321–328.

[12] X. Wu, Z. Li, and S. Chang, "Analyzing the harmonic structure in graph-based learning," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, 2013, pp. 3129–3137.

[13] ——, "New insights into laplacian similarity search," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015, pp. 1949–1957.

[14] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[15] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22Nd International Conference on Machine Learning*, ser. ICML '05, 2005, pp. 89–96.

[16] S. Agarwal, "Learning to rank on graphs," *Machine Learning*, vol. 81, no. 3, pp. 333–357, 2010.

[17] A. K. Menon and R. C. Williamson, "Bipartite ranking: A risk-theoretic perspective," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 6766–6867, 2016.

[18] S. P. Chepuri, S. Liu, G. Leus, and A. O. H. III, "Learning sparse graphs under smoothness prior," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2017, pp. 6508–6512.

[19] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "A latent variable model approach to pmi-based word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 385–399, 2016.

[20] X. Wu, Z. Li, A. M. So, J. Wright, and S. Chang, "Learning with partially absorbing random walks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, 2012, pp. 3086–3094.

[21] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, "Localization of the maximal entropy random walk," *Phys. Rev. Lett.*, vol. 102, p. 160602, 2009.

[22] J. G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," *IEEE Transactions on Cybernetics*, vol. 44, no. 9, pp. 1661–1672, 2014.

[23] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, pp. 39–43, 1953.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11, 2011.

[26] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[27] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia university image library (coil-20)," *Technical Report CUCS-005-96*, 1996.