JMLR: Workshop and Conference Proceedings 39:16–31, 2014 ACML 2014

Sparsity on Statistical Simplexes and Diversity in Social Ranking

Ke Sun KE.SUN@UNIGE.CH
Hisham Mohamed HISHAM.MOHAMED@UNIGE.CH
Stéphane Marchand-Maillet STEPHANE.MARCHAND-MAILLET@UNIGE.CH
Viper Group, Computer Vision and Multimedia Laboratory, University of Geneva

Editor: Dinh Phung and Hang Li

Abstract

We study sparsity on a statistical simplex consisting of all categorical distributions. This is different from the case in $\Re^m$ because such a simplex is a Riemannian manifold, a curved space. A learner with sparse constraints would be likely to fall into its low-dimensional boundaries. We present a novel analysis on the statistical simplex as a manifold with boundary. We investigate the learning dynamics in between high-dimensional models in the interior of the simplex and low-dimensional models on its boundaries. We study the differentiability of the cost function and its natural gradient with respect to the Riemannian structure.

We apply the proposed technique to social network analysis. Given a directed graph, the task is to rank a subset of influencer nodes. Here, sparsity means that the top-ranked nodes should present diversity in the sense of minimizing influence overlap. We present a ranking algorithm based on the natural gradient. It can scale up to graph datasets with millions of nodes. On real large networks, its top-ranked nodes are the most influential among several commonly-used techniques.

Keywords: Sparsity, Ranking, Information Geometry

1. Introduction

Sparsity has been a main topic of machine learning (Tibshirani, 1996; Ng, 2004; Zhao and Yu, 2006; Bach, 2008). The majority of previous works concentrated on studying sparsity in an Euclidean space, where a model parameter $\boldsymbol{\alpha} \in \Re^m$ is constrained to be likely on certain subspaces of $\Re^m$ through L_1-type regularization or, equivalently, a Laplace prior distribution of $\boldsymbol{\alpha}$. This results in a "simple" model, in the sense that only a few entries of $\boldsymbol{\alpha}$ are non-zero.

Recently, the notion of sparsity has been extended (Pilanci et al., 2012; Kyrillidis et al., 2012) to the statistical simplex [1] $\mathcal{S}^m = \left\{(\eta_1, \ldots, \eta_m) : \forall j, \eta_j \geq 0; \sum_{j=1}^m \eta_j \leq 1\right\}$, meaning that only a small number of η_j's are non-zero. L_1-norm-based techniques are not ideal because (1) L_1 norm depends on the coordinate system and thus is not an intrinsic measure; (2) L_1 norm in the η-coordinates already appears as a constraint. Pilanci et al. (2012) proposed a relaxation of the minimization problem on the number of non-zero η_j's. Kyrillidis et al. (2012) studied sparsity based on an Euclidean projection onto some sparse region

1. The upper script of a manifold, e.g. "m" in "$\mathcal{S}^m$", denotes the dimensionality.

on $\mathcal{S}^m$. In these methods, $\mathcal{S}^m$ is studied as a subset of the ambient R^{m+1}, and sparsity is derived from the Euclidean geometry. However, in many cases, $\boldsymbol{\eta} \in \mathcal{S}^m$ means a probability distribution. $\mathcal{S}^m$ is not Euclidean but instead has a unique information geometry (Rao, 1945; Čencov, 1982; Amari and Nagaoka, 2000). To underhand sparsity in such a geometric way and to study sparsity that is invariant under re-parametrization is of theoretical interest.

Based on existing methods, we present an information geometric analysis on the statistical simplex as a manifold with boundary. This is novel because past efforts mainly focused on the interior of $\mathcal{S}^m$. While a learner can jump in-between the boundary $\partial\mathcal{S}^m$ and inside $\mathcal{S}^m$ (Ghahramani and Beal, 2000; Xu, 2009), the learning dynamics near $\partial\mathcal{S}^m$ are not explicitly investigated. We discovered that the learning cost function is decomposed into a smooth term and a non-smooth term, where smoothness is defined on the manifold with boundary. The non-smooth term, as a coordinate-invariant regularization, helps to create singularities near $\partial\mathcal{S}^m$, where the gradient flow is always inward, i.e., from $\partial\mathcal{S}^m$ to the interior of $\mathcal{S}^m$.

As an applicative contribution, we investigate such sparsity in graph-based ranking (Page et al., 1999). The task is to rank a subset of nodes in a social network, so that they can maximally spread influence. This is reduced to inferring a probability distribution on the graph nodes. Sparsity in this context means that a limited number of nodes have a non-zero probability of being an influencer. This agrees with recent interests in graph-based information retrieval to retrieve a *diversity* of nodes (Zhu et al., 2007). We propose a scalable implementation along with a novel usage of natural gradient (Amari, 1998). Through experimenting on real large networks, we show that the proposed ranking most effectively discovers important nodes to maximally cover the network.

The rest of this paper is organized as follows. Section 2 introduces some prerequisites of information geometry, then presents an analysis on sparsity on statistical simplexes. The theoretical results (theorems 3 and 4) are in subsection 2.2. Section 3 discusses an application on social network ranking and the associated learning algorithm. Section 4 discusses related works and compares the proposed ranking with PageRank. Section 5 presents an experimental study on real large networks. Section 6 concludes and discusses possible extensions.

2. Sparsity on Statistical Simplexes

An observable random variable X, either discrete or continuous, is associated with a latent random binary vector $Y \in \left\{(y_1, \ldots, y_m) : \forall j, y_j = 0 \text{ or } 1; \sum_{j=1}^m y_j \leq 1\right\}$, with at most one bit equal to "1" and following a discrete distribution. We assume that $p(X \mid Y)$ is given. This simplified case lets us focus on the central issue, i.e. sparsity, and is useful for social network analysis (to be introduced in section 3). Based on a set of independent and identically distributed observations (i.i.d.) $\mathcal{X} = \{X^1, \ldots, X^n\}$, the problem is to infer a prior distribution $p(Y)$, where Y should be sparse. Sparsity means that only a small subset of bits in Y are activated, i.e., having a non-zero probability of being "1", while the rest bits are deactivated, i.e., always "0".

2.1. Prerequisites

$p(Y)$ is in the exponential family of distributions, as it can be written in the canonical form

$$p(Y \mid \boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{m} \theta_j y_j - \psi(\boldsymbol{\theta})\right), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ are the *canonical parameters* ranging in $\Re^m$, and $\psi(\boldsymbol{\theta}) = \log\left(1 + \sum_{j=1}^{m} \exp\theta_j\right)$. Another way of representing $p(Y)$ is by $p(Y \mid \boldsymbol{\eta}) = \sum_{j=1}^{m} y_j \eta_j + (1 - \sum_{j=1}^{m} y_j)\eta_0$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$ are the *expectation parameters* ranging in $\mathcal{S}^m - \partial\mathcal{S}^m$, and $\eta_0 = 1 - \sum_{j=1}^{m} \eta_j$. The θ- and η-coordinate systems relate with each other by the Legendre transformations (Amari and Nagaoka, 2000)

$$\frac{\partial \psi}{\partial \boldsymbol{\theta}} = \boldsymbol{\eta}, \quad \frac{\partial \varphi}{\partial \boldsymbol{\eta}} = \boldsymbol{\theta}, \tag{2}$$

where $\varphi(\boldsymbol{\eta}) = \sum_{j=0}^{m} \eta_j \log \eta_j$ is the negative entropy. Both $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ are strictly convex functions with respect to $\boldsymbol{\theta} \in \Re^m$ and $\boldsymbol{\eta} \in \mathcal{S}^m$, respectively.

The statistical manifold consisting of all such $p(Y)$ is equipped with a *Riemannian metric* (Lee, 2012), which can be intuitively understood as a local inner product defined on each point and varying smoothly along the manifold. It was showed (Rao, 1945; Čencov, 1982) that Fisher Information Metric (FIM) $g_{ij}(\boldsymbol{\theta}) = -E(\partial^2 \log p / \partial \boldsymbol{\theta}^2)$ is the unique Riemannian metric under some conditions, where $E(\cdot)$ is the expectation with respect to $p(Y \mid \boldsymbol{\theta})$. By eqs. (1) and (2), $g_{ij}(\boldsymbol{\theta}) = \partial^2\psi / \partial\boldsymbol{\theta}^2 = \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}$ coincides with the Jacobi matrix $\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}$ of the coordinate transformation $\boldsymbol{\theta} \to \boldsymbol{\eta}$. Similarly, FIM with respect to the η-coordinates is $g_{ij}(\boldsymbol{\eta}) = \partial^2\varphi/\partial\boldsymbol{\eta}^2 = \partial\boldsymbol{\theta}/\partial\boldsymbol{\eta}$, which is the inverse of $g_{ij}(\boldsymbol{\theta})$. It can be verified that $g(\boldsymbol{\theta})$ and $g(\boldsymbol{\eta})$ are essentially the same metric by showing $\langle \boldsymbol{a}\partial\boldsymbol{\theta}, \boldsymbol{b}\partial\boldsymbol{\theta}\rangle_{g(\boldsymbol{\theta})} = \langle \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}\boldsymbol{a}\partial\boldsymbol{\eta}, \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}\boldsymbol{b}\partial\boldsymbol{\eta}\rangle_{g(\boldsymbol{\eta})}$. $\boldsymbol{a}\partial\boldsymbol{\theta}$ denotes the *vector field* (Lee, 2012) $\sum_{j=1}^{m} a_j \partial\theta_j$, which can be understood as real vectors in local linearizations of the Riemannian manifold $\mathcal{S}^m$. $\langle\cdot,\cdot\rangle_{g(\boldsymbol{\theta})}$ denotes the inner product with respect to the Riemannian metric $g(\boldsymbol{\theta})$.

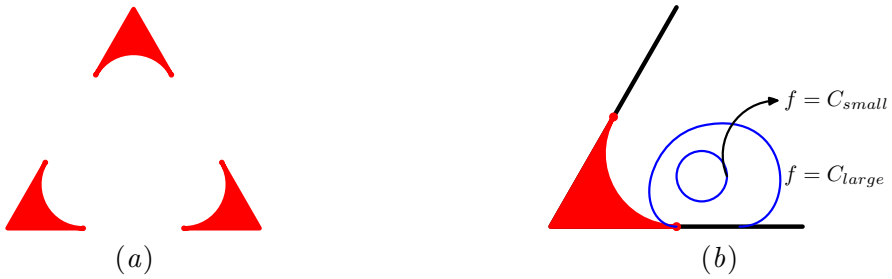
In this paper, FIM is used to compute the *natural gradient* (Amari and Nagaoka, 2000), i.e. the gradient with respect to the Riemannian geometry, of a smooth function f on $\mathcal{S}^m$. By definition, the natural gradient of f is $\texttt{grad}f = (g_{ij}(\boldsymbol{\theta}))^{-1}\partial f/\partial\boldsymbol{\theta} \cdot \partial\boldsymbol{\theta} = \partial f/\partial\boldsymbol{\eta} \cdot \partial\boldsymbol{\theta}$. $\texttt{grad}f$ is invariant to the choice of the coordinate system. For example, with respect to the η-coordinates, $\texttt{grad}f = (g_{ij}(\boldsymbol{\eta}))^{-1}\partial f/\partial\boldsymbol{\eta} \cdot \partial\boldsymbol{\eta} = \partial f/\partial\boldsymbol{\theta} \cdot \partial\boldsymbol{\eta}$ is exactly the same $\texttt{grad}f$ up to coordinate transformation.

Maximum-likelihood learning can be implemented (Amari, 1995) by

$$\begin{cases} \boldsymbol{\theta}^i = \boldsymbol{\theta} + \boldsymbol{c}^i, \quad \forall i = 1, \ldots, n; \\ \tilde{\boldsymbol{\eta}} = \sum_{i=1}^{n} \boldsymbol{\eta}^i / n; \\ \min_{\boldsymbol{\theta}} E_\tau(\boldsymbol{\theta}), \quad E_\tau(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}) + \tau\varphi(\tilde{\boldsymbol{\eta}}) - \boldsymbol{\theta}^T\tilde{\boldsymbol{\eta}}. \end{cases} \tag{3}$$

The first equation is the Bayes' rule, where $c^i_j = \log p(X^i \mid Y_j = 1) - \log p(X^i \mid Y = \mathbf{0})$, and $\boldsymbol{\theta}^i$ denotes the posterior estimation with regard to X^i [2]. The second equation summarizes the posterior estimations into a new $\tilde{\boldsymbol{\eta}}$ (or $\tilde{\boldsymbol{\theta}}$). The last line in eq. (3) minimizes the difference

2. One can re-write $\boldsymbol{\theta}^i = \boldsymbol{\theta} + \boldsymbol{c}^i$ in the canonical form of Bayes' rule in the η-coordinates.



(a) (b)

Figure 1: (a) The low-entropy region $\{\boldsymbol{\eta} : -\varphi(\boldsymbol{\eta}) \leq C\}$ in $\mathcal{S}^2$, where $C \approx 0.6$; (b) A constrained optimization $\min f(\boldsymbol{\eta})$, s.t. $-\varphi(\boldsymbol{\eta}) \leq C$. $f(\boldsymbol{\eta})$ is showed by the blue contours.

between $\boldsymbol{\theta}$ and its image $\tilde{\boldsymbol{\eta}}$ after the first and second equations, under the intuition that a locally optimal $\boldsymbol{\theta}$ should coincide with $\tilde{\boldsymbol{\eta}}$. Note, E_1 is exactly the Kullback-Leibler (KL) divergence from $\tilde{\boldsymbol{\eta}}$ to $\boldsymbol{\theta}$. This learning is supported by the following propositions.

Proposition 1 *$E_1(\boldsymbol{\theta}) \geq 0$; $E_1(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta}$ is a stationary point of the log-likelihood function $L(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(X^i \,|\, \boldsymbol{\theta})$.*

Proposition 1 says that $E_1(\boldsymbol{\theta})$ is an "indicator function", reaching its minimum at, and only at, the stationary points of $L(\boldsymbol{\theta})$. The proof is straightforward by writing $L(\boldsymbol{\theta}) = \sum_{i=1}^n \log \sum_Y \left(p(Y \,|\, \boldsymbol{\theta})\; p(X^i \,|\, Y)\right)$ and computing its differential according to eq. (1).

Proposition 2 *(1)* $\texttt{grad}\, L = n(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta})\partial\boldsymbol{\eta}$*; (2)* $\texttt{grad}\, E_\tau = \left(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}} + \frac{\partial\tilde{\boldsymbol{\eta}}}{\partial\boldsymbol{\theta}}(\tau\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\right)\partial\boldsymbol{\eta}$.

Proposition 2, which can be derived from the definition of the natural gradient introduced earlier and eqs. (2) and (3), gives the natural gradients of $L(\boldsymbol{\theta})$ and $E_\tau(\boldsymbol{\theta})$. It shows that minimizing $E_1(\boldsymbol{\theta})$ instead of maximizing $L(\boldsymbol{\theta})$ benefits from another gradient term pulling together $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$. This explains the faster convergence of learning as a two-body problem (Amari, 1995). Here, the two bodies $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\eta}}$, which are marginal distributions $p(Y)$, and the learning gradient flow are all in one simple space $\mathcal{S}^m$. Equation (3) is just one (representative) method in a widely-studied spectrum. Therefore, the information geometric analysis in the following subsection 2.2 could be useful in more general contexts.

2.2. Sparsity on Statistical Simplexes

The η-coordinates expose a hierarchy of statistical manifolds. This allows us study singular regions (Amari et al., 2006) and impose sparsity on Y. The closed simplex $\mathcal{S}^m$ is a *manifold with boundary* (Lee, 2012), where any point has a neighborhood which is like an open subset of $R_+^m = \{\boldsymbol{\alpha} \in \Re^m : \alpha_1 \geq 0\}$. There are "corners" in $\mathcal{S}^m$, which do not satisfy such a property. They are less interesting and will be ignored in subsequent discussions. A face of $\mathcal{S}^m$ is a statistical manifold with exactly the same structure as $\mathcal{S}^m$ but one less dimension, corresponding to the sparse case where some bit of Y is deactivated. This paper considers the learning dynamics on $\mathcal{S}^m$ as a whole without excluding $\partial\mathcal{S}^m$. A learner can go from inside $\mathcal{S}^m$ to $\partial\mathcal{S}^m$ or the other way round.

A natural idea to make Y sparse is to penalize the entropy $-\varphi$ by setting $\tau < 1$ in eq. (3). This gives an intrinsic regularization as the entropy is invariant to the choice of the coordinate system. This can be understood by the fact that $\varphi(\boldsymbol{\eta})$ is the KL-divergence from $\boldsymbol{\eta}$ to the simplex center plus some constant. To gain some intuitions, consider $\mathcal{S}^2$. Figure 1(a) shows the low entropy region $\{\boldsymbol{\eta} : -\varphi(\boldsymbol{\eta}) \leq C\}$, where $C \approx 0.6$. Such a region has some sharp corners, which can easily trap the optimizer of a constrained problem $\min f(\boldsymbol{\eta})$, s.t. $-\varphi(\boldsymbol{\eta}) \leq C$, where $f(\boldsymbol{\eta})$ is a smooth function on $\mathcal{S}^2$. This is only an intuitive view in the η-coordinate system. Formally, we have to consider the smoothness of the learning cost function E_τ in eq. (3). A *smooth function f on the manifold with boundary* $\mathcal{S}^m$ means that the differentials of f continuously extend to an open neighbourhood of $\boldsymbol{\eta}$ in $\Re^m$ for any $\boldsymbol{\eta} \in \mathcal{S}^m$. Even if $\boldsymbol{\eta}$ is a bit outside $\mathcal{S}^m$, these differentials are still well-defined. We have the following result [3].

Theorem 3 *Assume that $\forall j, \exists i,$ s.t. $p(X^i \,|\, Y_j = 1) > 0$. Then, (1) E_1 is a smooth function on $\mathcal{F} = \{\boldsymbol{\eta} \in \mathcal{S}^m \,|\, \forall i, p(X^i \,|\, \boldsymbol{\eta}) > 0\}$; (2) $\forall \tau < 1$, E_τ is continuous on $\mathcal{F}$ but non-differentiable on $\{\boldsymbol{\eta} \in \mathcal{F} : \exists j, \eta_j = 0; \forall i \neq j, \eta_i > 0\}$.*

In the above theorem 3, the assumption means that any bit in Y is associated to at least one observation X^i, otherwise it can be removed without affecting the system. $\mathcal{F}$ is a feasible region consisting of all such $\boldsymbol{\eta}$ that "covers" all observations. If there is some redundant bit(s) in Y that can be deactivated, then $\mathcal{F} \cap \partial\mathcal{S}^m \neq \emptyset$. The smoothness of E_1 does not rely on the choice of the coordinate system and therefore reflects an intrinsic property. A learner based on E_1 is "unaware" of $\partial\mathcal{S}^m$, meaning that it does not treat $\partial\mathcal{S}^m$ in a particular way. During learning, it could go from inside $\mathcal{S}^m$ to $\partial\mathcal{S}^m$ or the other way round. Often, it tends to go from $\partial\mathcal{S}^m$ to inside $\mathcal{S}^m$, because a model inside $\mathcal{S}^m$ has a higher complexity and a higher potential likelihood.

By making τ smaller than 1, E_τ becomes non-differentiable on $\partial\mathcal{S}^m$. This reveals an interesting relationship with sparsity on $\Re^m$ (Ng, 2004). L_1 norm, which is non-differentiable on $\{\boldsymbol{\alpha} \in \Re^m : \exists j, \alpha_j = 0\}$, is used to enforce sparsity on $\Re^m$. The entropy function $-\varphi$, which is non-differentiable on $\partial\mathcal{S}^m$, is used to enforce sparsity on $\mathcal{S}^m$. The level sets of φ have the form $d\varphi = \sum_{j=1}^m \partial\varphi/\partial\eta_j d\eta_j = \sum_{j=1}^m \theta_j d\eta_j = 0$. This differential representation is similar to the level sets of L_2 norm in an Euclidean space. This helps to understand why φ plays a similar role of a norm. Theorem 3 tells that the surface of E_τ is singular on $\partial\mathcal{S}^m$, but it does not describe its gradient flow near $\partial\mathcal{S}^m$ inside $\mathcal{S}^m$. This is covered by the following theorem.

Theorem 4 *If $\tau < 1$, $\langle \mathtt{grad} E_\tau, \partial/\partial\eta_j \rangle_g \to \infty$ as $\eta_j \to 0^+$ inside $\mathcal{F}$.*

Remark 5 *Although the statement is based on η-coordinates, the natural gradient $\mathtt{grad} E_\tau$ is invariant to the coordinate system. In another coordinate system ζ, $\mathtt{grad} E_\tau$ is still "inward", meaning it flows from $\partial\mathcal{S}^m$ to inside $\mathcal{S}^m$.*

As we are minimizing E_τ in eq. (3), learning is along the vector field $-\mathtt{grad} E_\tau$. Theorem 4 says that whenever the learner approaches $\partial\mathcal{S}^m$ from inside $\mathcal{S}^m$, and η_j becomes small enough, it will go to $\partial\mathcal{S}^m$. Therefore, there is a continuous "attractive region" near $\partial\mathcal{S}^m$,

3. See the appendix at http://cui.unige.ch/~sun/acml2014supp.pdf for the proofs.

where any $\boldsymbol{\eta}$ will be pulled into $\partial\mathcal{S}^m$. The size of this region depends on τ. If τ is slightly smaller than 1, the surface of E_τ is only bent down near $\partial\mathcal{S}^m$. As τ turns smaller, the attractive region widens. This also means that a learner can only reach $\partial\mathcal{S}^m$ from inside $\mathcal{S}^m$ but not go from $\partial\mathcal{S}^m$ to inside $\mathcal{S}^m$. From an algorithmic perspective, the problem scale is reduced during optimization. Learning becomes more and more efficient. On the other hand, any local optimum solution on $\partial\mathcal{S}^m$ also has such attraction to a leaner inside $\mathcal{S}^m$. If the learner falls into $\partial\mathcal{S}^m$ and deactivates a bit of Y that is in the global optimal solution, there is no way to reverse it in subsequent learning. The optimization must carefully explore the feasible region inside $\mathcal{S}^m$ before falling into $\partial\mathcal{S}^m$.

Consider E_τ as a function of $(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}) \in \mathcal{S}^m \times \mathcal{S}^m$ and the case $0 \leq \tau < 1$. By $\partial E_\tau / \partial \tilde{\boldsymbol{\eta}} = 0$, we get $\tau\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$. As the absolute value of θ is large near $\partial\mathcal{S}^m$, $\min E_\tau$ causes an isotropic scaling $\boldsymbol{\theta} \to \tilde{\boldsymbol{\eta}}$ towards $\partial\mathcal{S}^m$ in the θ-coordinate system. From a second-order view, the Hessian of E_τ is

$$H = \begin{bmatrix} g(\boldsymbol{\theta}) & -I \\ -I & \tau g(\tilde{\boldsymbol{\eta}}) \end{bmatrix}, \tag{4}$$

where g is FIM. E_τ is convex with respect to $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\eta}}$ individually. The joint convexity is guaranteed if and only if $\tau g(\tilde{\boldsymbol{\eta}}) \succeq g^{-1}(\boldsymbol{\theta})$, which is equivalent to $\tau g(\tilde{\boldsymbol{\eta}}) \succeq g(\boldsymbol{\eta})$. By simple derivations, $|g(\boldsymbol{\eta})|$ turns large as $\boldsymbol{\eta}$ moves from the simplex center to $\partial\mathcal{S}^m$. Therefore, the joint convexity of E_τ means that $\tilde{\boldsymbol{\eta}}$ should be closer to $\partial\mathcal{S}^m$ as compared to $\boldsymbol{\theta}$. This scaling in the θ-coordinates forms a mechanism, making the two-body system to be likely to reach $\partial\mathcal{S}^m$. Note, we do not consider the case $\tau < 0$, when E_τ is concave with respect to $\tilde{\boldsymbol{\eta}}$ and easily causes trivial solutions without strong constraints.

3. Application to Graph-based Ranking

Consider a social network given by a directed graph $\mathcal{G} = (\mathcal{V}; \mathcal{E})$, where $\mathcal{V}$ are the nodes, and $\mathcal{E} \subset \{(i, j) : i \in \mathcal{V}; j \in \mathcal{V}\}$ are the links. Each link $(i, j) \in \mathcal{E}$ from node i to node j is associated with a weight $w_{ij} > 0$. Usually, (i, j) means that j can influence i with the strength w_{ij}. For example, in twitter, (i, j) means that i reads micro-blogs posted by the individual j; in citation networks, (i, j) means that the article i is based on a previous article j. A subset $\mathcal{V}_I \subset \mathcal{V}$ of size $|\mathcal{V}_I| = m$, referred to as the *influencers*, are considered as potential candidates to emit influence. Their states of being chosen can be represented as an m-dimensional latent binary vector Y as in section 2. Without loss of generality, we set $\mathcal{V}_I$ to be the set of nodes with at least one incoming link. On the other hand, a target population $\mathcal{V}_O \subset \mathcal{V}$, characterized by the random variable X, plays the role of receiving influence. By default, we set $\mathcal{V}_O$ to be all nodes in $\mathcal{V}$ with at least one out-going link. An influencer in $\mathcal{V}_I$ indexed by j $(1 \leq j \leq m)$ can influence any node in $\mathcal{V}_O$ indexed by i $(1 \leq i \leq n)$ according to a given *influence matrix* $F_{m\times n}$. F gives the conditional distribution $p(X \,|\, Y)$ such that $\forall j, \forall i, F_{ji} \geq 0$ and $\forall j, \sum_{i=1}^n F_{ji} = 1$. F is usually sparse, pre-computed according to the graph structure and the link weights. We let

$$F_{ji} = \begin{cases} w_{ij} / \sum_{i:i\to j} w_{ij} & \text{if } (i, j) \in \mathcal{E}; \\ 0 & \text{if otherwise,} \end{cases} \tag{5}$$

meaning that an influencer j influences each of its predecessors with a strength proportional to the link weight. Based on the modeling in section 2, a prior distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$

can be learned, where $\forall j$, η_j is the probability of activating the influencer j. This $\boldsymbol{\eta}$ can be used to rank the influencers while presenting diversity in the ranking results.

Ranking Consider an information diffusion process in a social network $\mathcal{G}$. A piece of information, e.g. marketing material, is first distributed to an influencer j then passed to the network according to $F_{j\bullet}$. In such a two-step scheme, maximizing the log-likelihood $\sum_{i=1}^{n} \log p(X^i \,|\, \boldsymbol{\eta})$ means to maximize the influence coverage. The maximum likelihood solution means an optimal scheme to allocate the information source. It tells that some influencers with large weights are preferred among the others in distributing information. By eq. (3), $\boldsymbol{\eta}^i$ means that given an influenced node X^i, how likely such influence comes from each influencer. Therefore, $\tilde{\boldsymbol{\eta}}$ means the percentage of actual influenced nodes in $\{X^1, \ldots, X^n\}$ by each influencer. Minimizing E_1 in eq. (3) means that the random influencer should be placed according to its effective influence.

Diversifying The objective of diversification corresponds to sparsity of Y as discussed in section 2. Making Y sparse, i.e. making certain influencers deactivated, helps to save resources in real world applications. For example, only a limited number of marketing personnels have to be deployed for broadcasting a piece of information.

In the following, we study several simple cases of the proposed optimization, so that one can better understand the result ranking.

Proposition 6 *Consider the influencers* $\mathcal{V}_I = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \cdots$. $\forall \imath \neq \jmath$, $\texttt{pred}(\mathcal{V}_\imath) \cap \texttt{pred}(\mathcal{V}_\jmath) = \emptyset$, *where* $\texttt{pred}(\cdot)$ *is the set of predecessors. Denote the optimal solution of eq. (3) on the whole graph as* $\boldsymbol{\eta}^\star$. *Denote the optimal solution on the sub-graph induced by* $\mathcal{V}_l \cup \texttt{pred}(\mathcal{V}_l)$ *as* $\boldsymbol{\eta}^\star_l$. *Then* $\forall \tau \geq 0$, $\forall j \in \mathcal{V}_l$, $\eta^\star_j = \eta^\star_{lj} \, |\texttt{pred}(\mathcal{V}_l)|/|\texttt{pred}(\mathcal{V}_I)|$.

The condition in proposition 6 means that $\mathcal{G}$ can be partitioned into sub-graphs so that any two sub-graphs do not share a common influencer. Proposition 6 says that the distribution of $\eta^\star_j$ among the sub-graphs is proportional to the size of the target population. A random influencer is more likely in populated regions. In a sub-graph, if the number of influencers is large but the target population is small, there will be more competitions among the influencers, in the sense that only a small percentage of influencers will be activated. The optimal $\boldsymbol{\eta}^\star$ can be obtained by individually solving eq. (3) on the sub-graphs and assembling them based on proposition 6.

Proposition 7 $\forall j$, $|\texttt{upred}(j)|/|\mathcal{V}_O| \leq \tilde{\eta}_j \leq |\texttt{pred}(j)|/|\mathcal{V}_O|$, *where* $\texttt{upred}(j)$ *means the isolated predecessors of* j *having only* j *as their successor.*

Proposition 7 gives an upper bound and a lower bound of $\tilde{\boldsymbol{\eta}}$. Consider that $\boldsymbol{\eta}^\star$ and $\tilde{\boldsymbol{\eta}}^\star$ should be similar, this gives an estimated range on the resulting $\boldsymbol{\eta}^\star$. An influencer with at least one isolated predecessor cannot be deactivated, because its isolated predecessor(s) can only be influenced by it. An influencer j with many isolated predecessors is likely to have a large value of $\eta^\star_j$ and a high rank.

3.1. Implementation

We implemented an optimizer for eq. (3), focusing on the case $\tau = 0$ with very efficient optimization. This is because E_0 has the simplest expression and gives the most sparse solution, which is preferred in our social analysis. We call this ranking method "diversify".

The optimizer is based on the natural gradient (Amari, 1998). Recall from section 2 that the natural gradient of a function is a vector field that is invariant to the choice of the coordinate system. We choose the *spherical coordinates* $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)$, so that $\forall j = 0, \ldots, m, \beta_j = \sqrt{\eta_j}$. Lebanon (2003) proved that FIM is "equivalent" to the embedded Euclidean geometry on the hyper-sphere $\{\boldsymbol{\beta}\}$. We further have the following propositions.

Proposition 8 *For any smooth function f on $\mathcal{S}^m$,*

$$\texttt{grad}f = \frac{1}{4}(I - \boldsymbol{\beta}\boldsymbol{\beta}^T)\frac{\partial f}{\partial \boldsymbol{\beta}} \cdot \partial\boldsymbol{\beta} = \frac{1}{2}\sum_{j=0}^{m} \beta_j \left(\frac{\partial f}{\partial \eta_j} - \sum_{j=0}^{m} \eta_j \frac{\partial f}{\partial \eta_j}\right) \partial\beta_j, \tag{6}$$

where I is the identity matrix.

Proposition 9 $\forall j$, $\partial E_0/\partial\eta_j = -\sum_{i:i\to j} \eta_j^i (1 + \log\eta_j - \sum_{j=0}^m \eta_j^i \log\eta_j)/(n\eta_j)$.

Equation (6) is exactly the projected gradient on the hypersphere $\{\boldsymbol{\beta}\}$ up to constant scaling. Proposition 8 gives an easy way to apply and understand natural gradient. To optimize any cost function $f(\eta_0, \ldots, \eta_m)$ on $\mathcal{S}^m$, one can regard $\eta_0, \ldots, \eta_m$ as independent variables, compute $\partial f/\partial\eta_j$, $\forall j = 0, \ldots, m$, and then compute the natural gradient by proposition 8.

First, $\boldsymbol{\beta}^0$ is randomly initialized, so that the corresponding $\boldsymbol{\eta}^0$ is roughly uniform for all influencers. Then we update $\boldsymbol{\beta}^{t+1}$ $(t = 0, 1, \ldots)$ until convergence following the rule

$$\boldsymbol{\beta}^{t+1} \leftarrow \frac{\boldsymbol{\beta}^t - \gamma \texttt{grad}E_0}{\|\boldsymbol{\beta}^t - \gamma \texttt{grad}E_0\|},$$

where $\gamma > 0$ is a small learning rate, $\|\cdot\|$ is 2-norm, and $\texttt{grad}E_0$ is given by propositions 8 and 9. An intuitive explanation of the learning process is in subsection 4.2.

Despite that the proposed optimization is non-convex and hence does not guarantee a global optimum solution, we find that its convergence is quite fast. To get a rough idea, fig. 2 shows the evolution of E_0 in the first 100 iterations on a DBLP collaboration network and an autonomous system network. In both cases, the convergence is reached in ∼30 iterations. Such fast convergence guarantees scalability. By our `C++` implementation, to compute $\boldsymbol{\beta}$ on a graph of ∼1.5 million nodes only costs minutes on a normal PC.

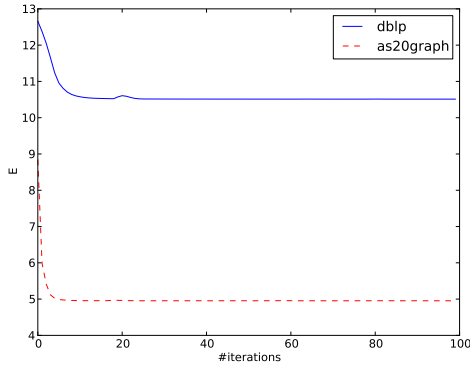


Figure 2: E_0 against the number of iterations

4. Discussion on Related Works

4.1. Bayesian methods

From a Bayesian perspective (Bishop, 1995), the i.i.d. observations $\{X^i\}$ induce on $\mathcal{S}^m$ a posterior distribution $p(\boldsymbol{\eta}\,|\,\{X^i\}) \propto \prod_i \left(\sum_Y p(X^i\,|\,Y)p(Y\,|\,\boldsymbol{\eta})\right) p(\boldsymbol{\eta})$. It is natural and interesting to view such a Bayesian inference on $\mathcal{S}^m$ as a manifold with boundary.

Consider the scenario to find an optimal $\boldsymbol{\eta}^\star \in \mathcal{S}^m$. The log-likelihood $L = \sum_i \log(\sum_Y p(X^i\,|\,Y)p(Y\,|\,\boldsymbol{\eta}))$, or a learning cost function derived from L, is in general smooth on a feasible region $\mathcal{F} \subset \mathcal{S}^m$, meaning that deactivating some bits in Y could bring down the value of L but does not create singularities. The learning dynamics near $\partial\mathcal{S}^m$ largely depend on the prior distribution $p(\boldsymbol{\eta})$. Traditional maximum likelihood learning uses flat priors, i.e. $p_U(\boldsymbol{\eta}) \propto 1$, resulting in a learner who is unaware of $\partial\mathcal{S}^m$. This is shown by the smoothness of E_1 in theorem 3. Jeffrey's prior (1946) is proportional to the Riemannian volume element (Lee, 2012) such that $p_J(\boldsymbol{\eta}) \propto |g(\boldsymbol{\eta})|^{1/2}$. It is non-informative, treating different points on $\mathcal{S}^m$ intrinsically equally. We give without proof that $p_J(\boldsymbol{\eta}) \propto \prod_{j=0}^m \eta_j^{-1/2}$. This leads to an inward flow that is similar to theorem 4 due to the non-smoothness of $\log t$ at $t = 0$. Such a similarity could partially justify the setting $\tau < 1$ used in this paper. Because $p_J(\boldsymbol{\eta}) \to \infty$ as $\boldsymbol{\eta} \to \partial\mathcal{S}^m$, $p_J(\boldsymbol{\eta})$, as well as $p(\boldsymbol{\eta}\,|\,\{X^i\})$, is not continuous on $\mathcal{S}^m$. It is easy to see that the prior used in this paper is $p_D(\boldsymbol{\eta}) \propto \prod_{j=0}^m \eta_j^{(1-\tau)\eta_j}$ [4]. Similar to $p_J(\boldsymbol{\eta})$, it has a concave shape on $\mathcal{S}^m$. The difference is that it dampens $p_J(\boldsymbol{\eta})$ near $\partial\mathcal{S}^m$ with a finite value on $\partial\mathcal{S}^m$, meaning that sparsity instead of small values of η_j is preferred. It yields a continuous $p(\boldsymbol{\eta}\,|\,\{X^i\})$ on $\mathcal{S}^m$, which is elegant in theory and establishes a global Bayesian view on $\mathcal{S}^m$.

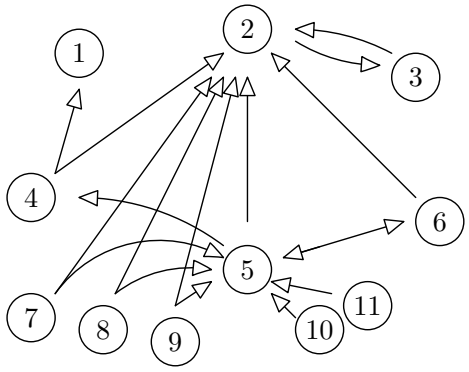
The model $\mathcal{S}^m$ can be assessed by evaluating $p(\{X^i\}) = \int_{\boldsymbol{\eta}\in\mathcal{S}^m} p(\{X_i\}\,|\,\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\eta}$. If $p_J(\boldsymbol{\eta})$ is used, the boundary regions occupy a large percentage of the total volume. For example, the volume of the region $\{\boldsymbol{\eta} \in \mathcal{S}^m : \exists j, \eta_j < 0.05 \text{ or } \eta_j > 0.95\}$ is at least $(1-0.9^m)$ times the total volume of $\mathcal{S}^m$. $p_J(\boldsymbol{\eta})$ as a non-informative prior emphasizes too much on such regions. As a result, the model assessment is largely based on such $\boldsymbol{\eta}$ with many small non-zero values of η_j's. This can be understood as a curse of dimensionality (Bellman, 1957) on the parameter manifold. $p_D(\boldsymbol{\eta})$, as a *weakly informative prior*, puts focus on the center of $\mathcal{S}^m$ with sufficient large η_j's. It could favor a simple model $\mathcal{S}^m$ over a complex model $\mathcal{S}^{m+1}$, if the maximum likelihood solution on $\mathcal{S}^{m+1}$ is near $\partial\mathcal{S}^{m+1}$, while the maximum likelihood solution on $\mathcal{S}^m$ is near the center. This essentially agrees with the idea of sparsity.

This work is connected to previous studies on singularities on statistical manifolds, where FIM is not well-defined (Amari et al., 2006; Cousseau et al., 2008; Park and Ozeki, 2009). Interestingly, we demonstrated that singularities can be helpful to learning, which is in contrast to the cases where singularities make learning difficult (Amari et al., 2006).

4.2. Diversified Ranking in Information Retrieval

Several proposals have been made for modeling and encouraging diversity in ranking. An extensive review of these works is proposed in (Raman et al., 2013). In summary, diversity is encouraged in ranked lists by modeling novelty and performing re-ranking (Carbonell and

4. Actually p_D is imposed on $\tilde{\boldsymbol{\eta}}$ instead of $\boldsymbol{\eta}$. We omit such a difference as $\tilde{\boldsymbol{\eta}}$, just like $\boldsymbol{\eta}$, is one of the two bodies in the learning dynamics.



Node	InDegree	PageRank	Diversify
1	1	3.3%	0
2	7	38.4%	68.1%
3	1	34.3%	10.0%
4	1	3.9%	0
5	6	8.1%	21.9%
6	1	3.9%	0
7	0	1.6%	0
8	0	1.6%	0
9	0	1.6%	0
10	0	1.6%	0
11	0	1.6%	0

Figure 3: A toy network from http://en.wikipedia.org/wiki/PageRank

Goldstein, 1998) following the Cascade Model of user behavior (Clarke et al., 2011). Query reformulation (Santos et al., 2010) also goes in the line of document re-ranking. Diversity may be associated with the inclusion of risk in the document ranking process. Risk may be viewed as ranking high novel but less-relevant documents (Wang and Zhu, 2009) or from the point of view of satisfying user intent (Agrawal et al., 2009). Alternatively, diversity may be embedded into the process of learning-to-rank by maximizing the expected user satisfaction over probable rankings (Radlinski et al., 2008; Slivkins et al., 2013).

PageRank

The widely-applied PageRank (Page et al., 1999) can be formulated by replacing the first two equations in eq. (3) with $\tilde{\boldsymbol{\eta}} = A^T\boldsymbol{\eta}$, where $A_{ij} = (1-\nu)/n + \nu\delta_{ij}/\mathtt{deg}(i)$, $\delta_{ij} = 1$ if $(i,j) \in \mathcal{E}$ and $\delta_{ij} = 0$ if otherwise, and $\nu = 0.85$ is a damping parameter. In this case, the problem can be solved with fixed point iterations $\boldsymbol{\eta} \leftarrow \tilde{\boldsymbol{\eta}}$. Comparatively, an updating scheme [5] of $\boldsymbol{\eta}$ based on proposition 2 is given by

$$\eta_j \leftarrow (1-\gamma)\eta_j + \frac{\gamma}{n}\sum_{i=1}^{n}\left(1+\theta_j - \boldsymbol{\theta}^T\boldsymbol{\eta}^i\right)\eta_j^i, \tag{7}$$

where γ is a learning rate. They both can be understood as a *voting process*. In PageRank, each node j votes for its successors uniformly, weighted by the current α_j. The amount of votes received by each node j determines the new value of α_j. In eq. (7), the voting is neither uniform nor strictly positive. For each predecessors i of node j, if θ_j is smaller than the threshold $(\boldsymbol{\theta}^T\boldsymbol{\eta}^i - 1)$, then node i casts a *negative vote* to node j. In this way, i chooses strong candidates from its successors and penalizes weak candidates. A compact list of candidates can be elected. In the toy example in fig. 3, diversify only selects three influencers, while most nodes are deactivated. For example, node 6 in the graph receives zero weight, because its predecessor node 5 is already influenced by node 2.

5. This is only an intuitive view. The learning is in the β-coordinates as explained in section 3.1.

Table 1: SNAP datasets used in the experiments

Dataset	#nodes	#edges	Directed	description
p2p-Gnutella04	10,876	39,994	Yes	Gnutella peer to peer network
p2p-Gnutella05	8,846	31,839	Yes	Gnutella peer to peer network
p2p-Gnutella06	8,717	31,525	Yes	Gnutella peer to peer network
web-BerkStan	685,230	7,600,595	Yes	Web graph of Berkeley and Stanford
soc-Pokec	1,632,803	30,622,564	Yes	Pokec online social network
cit-Patents	3,774,768	16,518,948	Yes	Citation network among US Patents
com-DBLP	317,080	1,049,866	No	DBLP collaboration network
com-Amazon	334,863	925,872	No	Amazon product co-purchase network
com-Youtube	1,134,890	2,987,624	No	Youtube online social network

5. Experiments

In this section, we evaluate the proposed ranking algorithm on real data. We select several medium-to-large networks from Stanford large network dataset collection (SNAP)[6]. The datasets used cover various domains as shown in table 1.

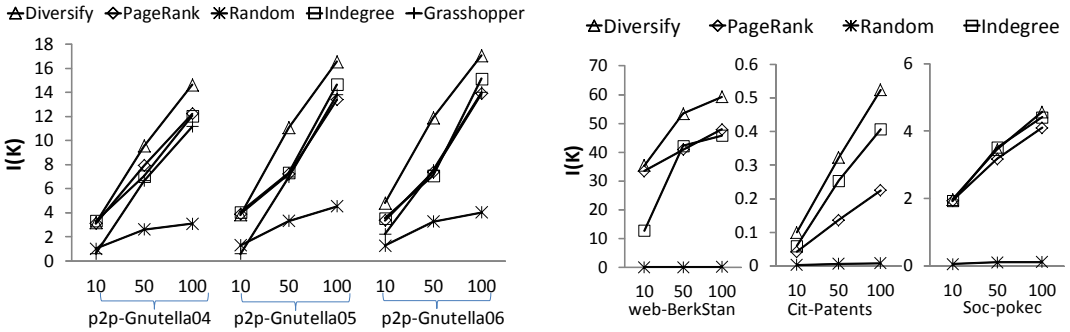
5.1. Spread Information

We investigate how the ranking approaches can influence the network by spreading information. We select 5 different algorithms: `Random` (random selection), `Indegree` (ranking by indegree), `Grasshopper` (a graph-based ranking algorithm achieving both diversity and centrality (Zhu et al., 2007)), `PageRank` (Page et al., 1999) and `Diversify`. For each algorithm on each dataset, we extract the top-ranked nodes (seeds) and count the number of nodes that can be influenced by them, that is, the number of nodes that link to these seeds. This performance measurement makes sense in numerous applications, such as finding valuable nodes for content distribution networks (CDN), or finding impactive patents in patent-citation networks. Simply picking the mostly-linked nodes is likely to fail in such tasks, because a pair of well-connected nodes often have high-overlap in their influence.

Figure 4 shows the number of influenced nodes against the number of seeds. It is clear that the proposed algorithm most effectively covers the network, followed by `Grasshopper`, `PageRank` and `Indegree`. `Random` performs the worst as expected. `Grapsshopper` selects nodes in a greedy manner. To select each node requires a large matrix inversion. It fails to operate in reasonable time on large datasets with millions of nodes, and thus no corresponding result is shown.

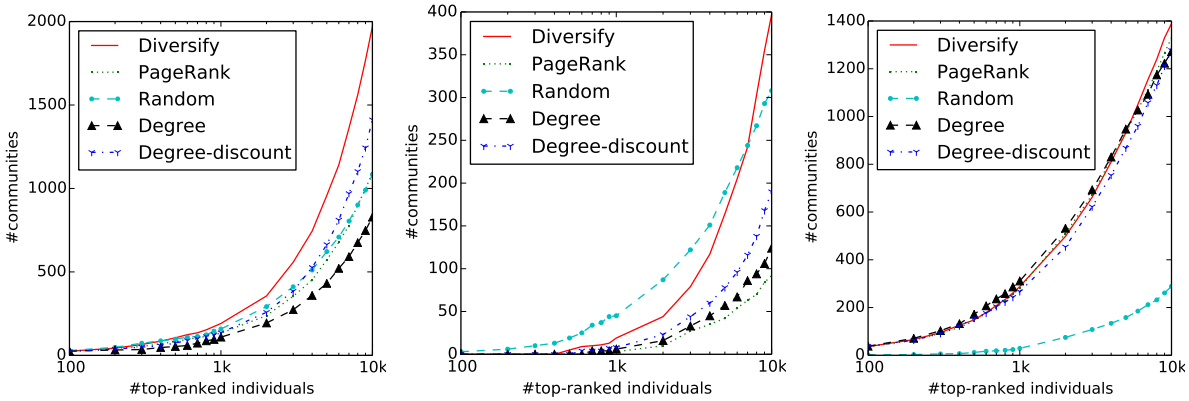
The good performance of `Diversify` as compared to `Grasshopper` that is designed for a similar purpose is explained as follows. `Diversify` is capable of global coordination: an influencer with large indegree but bad cooperation can be kicked out. `Grasshopper` is a greedy algorithm. If a bad influencer is already selected, there is no way to reverse it. `Diversify` performs better in network coverage and computation efficiency.

6. https://snap.stanford.edu/data



(*a*) Small Networks (*b*) Large Networks

Figure 4: The percentage (y-axis) of unique nodes that link to the top-k nodes with respect to k (x-axis)



(*a*) DBLP (*b*) Amazon (*c*) Youtube

Figure 5: The number of ground-truth communities (y-axis) (among the top 5000 communities on different social networks) covered by the top-k-ranked nodes against k (x-axis), with k ranging from 100 to 10,000

5.2. Community Coverage in Social Networks

Consider social networks where the users form communities. In many applications, the goal of ranking is to identify representative individuals (Chen et al., 2009; Sun et al., 2013), who are both well-connected individually, and involve in diverse communities. For example, in scientific collaboration networks, it is a common task to select researchers in distinct sub-areas for organizing events. We use datasets with ground-truth communities and keep these ground-truth from the ranking algorithms. We evaluate the ranking results by counting the number of communities covered by the top-ranked nodes.

We select 3 social networks: 1) `DBLP` computer science collaboration network, where the links represent co-authorships between authors, and the communities are defined by publication venue, etc. 2) `Amazon` product network, where the links represent frequently-occurred co-purchases between products, and the communities are defined by product categories.

3) `Youtube` social network, where the links are user friendships, and the communities are user-created groups on this platform. For each dataset, 5000 high quality ground-truth communities are known ahead (Yang and Leskovec, 2012). Unfortunately, we did not find any publicly-available directed graph with ground-truth community information. The proposed approach can handle these undirected graphs by replacing each edge with two opposite-directed links. We remove `Grasshopper` from the comparison, because it does not scale well on large datasets as shown in section 5.1. We compare another technique, `Degree-discount` (Chen et al., 2009), a simple heuristic for undirected graphs, which can effectively select nodes to maximize their influence spread.

Figure 5 shows the number of communities covered by the top-k-ranked nodes against k, where k ranges from 100 to 10,000. We assume that for even larger values of k, the community coverage is less interesting, because selecting more seeds usually implies costing more resources. On small values of k, all methods perform similarly. The difference shows up as k increases. We see that in general `Diversify` is among the top methods that cover the largest number of communities. `Degree-discount` also achieves good performance as compared to `PageRank` and `Degree` on the first two datasets. The good performance of `Random` on the Amazon dataset is because the product-categories are small-size communities which are likely to be disjoint. A large co-purchase number does not guarantee high coverage of such communities. This is different from social networks, where popular individuals tend to belong to more communities. However, the seed quality by `Random` is expected to be lower than `Diversify`, because `Diversify` considers the link structure and selects nodes with high degrees. On `Youtube`, the community distribution is much more sparse. Around 90% individuals are not signed up in any communities. If k is below 5,000, counting the most active individuals by `Indegree` effectively covers different communities. On the range from $5,000$ to $10,000$, `Diversify` still performs best (note that x-axis is log-scale).

5.3. Graph-based Movie Ranking

To apply the proposed method to a real-world ranking scenario, we select the `MovieLens` dataset[7] consisting of ∼10 million 5-star-ratings from ∼72,000 users to ∼10,000 movies. We check whether each user gives at least 4.5 stars to each movie, resulting in a directed user-rate-movie graph with 78,377 nodes and 2,129,834 edges. We also check whether each pair of movies are simultaneously rated higher than 4.5 by at least 10 users. The largest connected component gives an undirected movie-co-like graph with 5,316 nodes and 1,594,531 edges.

Table 2 shows the top-ranked movies based on the co-like graph [8]. The top-15 by `PageRank` concentrated on movies in the 1990s. There are two episodes of "Star Wars", which are similar in contents. `Diversify` presents a wider range in the sense of release-time. It could therefore satisfy more users. It discovers non-English movies like "Wooden Man's Bridge". An interesting observation is that it selects more old classical movies like "Citizen Kane". Such an observation is consistent when we vary the settings such as the threshold of simultaneous co-likes (which is set to 10 in this experiment) used to construct the co-like graph. Table 3 shows the coverage of the top-ranked movies. Movie coverage is based on the number of movies directly linked to the top list in the movie-co-like graph.

7. The 10M dataset at http://grouplens.org/datasets/movielens/ is used.
8. See http://imdb.com for related information of the movies.

Table 2: The top-15 movies on the `MovieLens` dataset based on co-like relationships. The unique movies discovered by `Diversify` are displayed in bold.

Rank	`PageRank`	`Diversify`
1	Pulp Fiction (1994)	Pulp Fiction (1994)
2	Shawshank Redemption (1994)	Star Wars IV - A New Hope (1977)
3	Matrix (1999)	Shawshank Redemption (1994)
4	Godfather (1972)	Godfather (1972)
5	Star Wars IV - A New Hope (1977)	Matrix (1999)
6	Silence of the Lambs (1991)	**Secret Agent (1996)**
7	American Beauty (1999)	**Wooden Man's Bride (1994)**
8	Fargo (1996)	Forrest Gump (1994)
9	Forrest Gump (1994)	Fargo (1996)
10	Raiders of the Lost Ark (1981)	**Citizen Kane (1941)**
11	Sixth Sense (1999)	**For the Moment (1994)**
12	Schindler's List (1993)	**Lord of the Rings: The Two Towers (2002)**
13	Star Wars V (1980)	Sixth Sense (1999)
14	Usual Suspects (1995)	**Dr. Strangelove (1964)**
15	Fight Club (1999)	American Beauty (1999)

Table 3: Movie coverage among 5,316 movies and user coverage among 68,860 users by the top-ranked movies. The "Sparsity" column shows the percentage of movies weighted greater than 10^{-7} based on the rankings in the co-like graph.

	Movie Coverage		User Coverage		Sparsity
	top-10	top-100	top-10	top-100	
`PageRank`	5055	5236	50497	65639	100%
`Diversify`	5058	5301	52042	66182	2.45%

User coverage is based on the number of users linked to the top list in the user-rate-movie graph. In both cases, `Diversity` is able to cover more movies or users. Unlike `PageRank`, its ranking is sparse, meaning that the majority movies receive a score of zero. This could be useful in scenarios such as purchasing a small number of representative movies.

6. Conclusion

We present an information geometric analysis on the sparsity on $\mathcal{S}^m$. The discovery of the inward flow near $\partial\mathcal{S}^m$ helps to understand the learning dynamics and the model variation. We advocate a weakly informative prior derived from the negative entropy function. It is continuous on $\mathcal{S}^m$ and is meaningful in Bayesian inference and model selection. We apply the proposed sparsity technique on graph-based ranking problems to enforce diversity. It scales to social graphs with tens of millions of nodes. In our experiments, the proposed method most effectively covers social networks among several commonly-used techniques.

Acknowledgments

This work is supported by the European COST Action on Multilingual and Multifaceted Interactive Information Access (MUMIA) via the Swiss State Secretariat for Education and Research (SER grant C11.0043).

References

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *International Conference on Web Search and Data Mining (WSDM)*, pages 5–14, 2009.

S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

S. Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998.

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. AMS and OUP, 2000. (Published in Japanese in 1993).

S. Amari, H. Park, and T. Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural Comput.*, 18(5):1007–1065, 2006.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *NIPS 21*, pages 105–112. 2008.

R. E. Bellman. *Dynamic Programming*. PUP, Princeton, NJ, USA, 1957.

C. M. Bishop. *Neural Networks for Pattern Recognition*. OUP, New York, NY, USA, 1995.

J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. AMS, 1982. (Published in Russian in 1972).

W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *SIGKDD*, pages 199–208, 2009.

C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *International Conference on Web Search and Data Mining (WSDM)*, pages 75–84, 2011.

F. Cousseau, T. Ozeki, and S. Amari. Dynamics of learning in multilayer perceptrons near singularities. *IEEE Trans. Neural Netw.*, 19(8):1313–28, 2008.

Zoubin Ghahramani and Matthew J. Beal. Variational inference for bayesian mixtures of factor analysers. In *NIPS 12*, pages 449–455. 2000.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A*, 186(1007):453–461, 1946.

A. Kyrillidis, S. Becker, V. Cevher, and C. Koch. Sparse projections onto the simplex. *CoRR*, abs/1206.1529, 2012.

G. Lebanon. Learning Riemannian metrics. In *UAI*, pages 362–369, 2003.

J. M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, USA, 2nd edition, 2012.

A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *ICML*, pages 78–86, 2004.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.

H. Park and T. Ozeki. Singularity and slow convergence of the EM algorithm for Gaussian mixtures. *Neural Process. Lett.*, 29(1):45–59, 2009.

M. Pilanci, L. E. Ghaoui, and V. Chandrasekaran. Recovery of sparse probability measures via convex programming. In *NIPS 25*, pages 2420–2428. 2012.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, pages 784–791, 2008.

K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: exploring intrinsic diversity in web search. In *SIGIR*, pages 463–472, 2013.

C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Cal. Math. Soc.*, 37(3):81–91, 1945.

R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *International Conference on World Wide Web (WWW)*, pages 881–890, 2010.

A. Slivkins, F. Radlinski, and S. Gollapudi. Ranked bandits in metric spaces: Learning diverse rankings over large document collections. *J. Mach. Learn. Res.*, 14(1):399–436, 2013.

K. Sun, D. Morrison, E. Bruno, and S. Marchand-Maillet. Learning representative nodes in social networks. In *PAKDD*, pages 25–36, 2013.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, 58(1):267–288, 1996.

J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR*, pages 115–122, 2009.

L. Xu. Learning algorithms for RBF functions and subspace based functions. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, pages 60–94. IGI Global, 2009.

J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *CoRR*, abs/1205.6233, 2012.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7(Nov):2541–2563, 2006.

X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, pages 97–104, 2007.