

---

# The Product Cut

## (Supplementary Material)

---

**Xavier Bresson**

Nanyang Technological University  
Singapore  
xavier.bresson@ntu.edu.sg

**Thomas Laurent**

Loyola Marymount University  
Los Angeles  
tlaurent@lmu.edu

**Arthur Szlam**

Facebook AI Research  
New York  
aszlam@fb.com

**James H. von Brecht**

California State University, Long Beach  
Long Beach  
james.vonbrecht@csulb.edu

### 1 Algorithmic Details: Closed-form Solution of the Linear Program (16)

The algorithm for the Product Cut objective proceeds by solving a sequence of random linear programs

$$\text{maximize } L_k(F) \tag{1}$$

$$\text{subject to } F \in C \tag{2}$$

$$\psi_i(F) = 0 \text{ for } i \in \mathcal{I}_k. \tag{3}$$

where  $L_k(F)$  is the linearization of the energy  $\mathcal{E}(F)$  around the current iterate  $F^k$ . This LP has closed-form solution obtained by **gradient thresholding**. Starting from (P-rlx), we see that the convex function  $\mathcal{E}$  to be maximized is

$$\mathcal{E}(F) = \mathcal{E}(f_1, \dots, f_R) = \sum_{r=1}^R e(f_r) \tag{4}$$

and its linearization around the current iterate  $F^k = (f_1^k, \dots, f_R^k)$  is

$$L_k(F) = L_k(f_1, \dots, f_R) = \mathcal{E}(f_1^k, \dots, f_R^k) + \sum_{r=1}^R \langle \nabla e(f_r^k), f_r - f_r^k \rangle \tag{5}$$

Recall that  $C$  is the bounded convex set  $[0, 1]^n \times \dots \times [0, 1]^n$  and the  $n$  affine constraints  $\psi_i(F) = 0$  correspond to the row-stochastic constraints  $\sum_{r=1}^R f_{i,r} = 1$ . Plugging (5) in the Linear Program (1)-(3) and ignoring the constant terms, we obtain:

$$\text{maximize } \sum_{r=1}^R \langle \nabla e(f_r^k), f_r \rangle = \sum_{i=1}^n \sum_{r=1}^R f_{ir} \nabla e(f_r^k)_i \tag{6}$$

$$\text{subject to } 0 \leq f_{i,r} \leq 1 \text{ for } 1 \leq i \leq n \text{ and } 1 \leq r \leq R \tag{7}$$

$$\sum_{r=1}^R f_{i,r} = 1 \text{ for } i \in \mathcal{I}_k \tag{8}$$

where  $\nabla e(f_r^k)_i$  stands for the  $i^{\text{th}}$  entry of the vector  $\nabla e(f_r^k)$ . Note that the above problem decouples in  $i$  and can be solved explicitly: if  $i \in \mathcal{I}_k$  then

$$f_{i,r}^{k+1} = \begin{cases} 1 & \text{if } \nabla e(f_r^k)_i > \nabla e(f_s^k)_i \text{ for all } s \neq r \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

---

**Algorithm 1** Gradient Thresholding Algorithm to solve the Linear Program (1)-(3)

---

**Input:** Current iterate  $F^k = (f_1^k, \dots, f_R^k)$  and set  $\mathcal{I}_k \subset \{1, \dots, n\}$ .

```

for  $r = 1$  to  $R$  do
   $\hat{f}_r = f_r^k / (\sum_{i=1}^n f_{i,r}^k)$ 
  Solve  $M_\alpha u_r = \hat{f}_r$ 
   $g_{i,r} = \hat{f}_{i,r} / u_{i,r}$  for  $i = 1, \dots, n$ .
  Solve  $M_\alpha^T v_r = g_r$ 
   $h_r = \log u_r + v_r - \mathbf{1}$ 
end for
for all  $i \in \mathcal{I}_k$  do
   $f_{i,r}^{k+1} = \begin{cases} 1 & \text{if } r = \arg \max_s h_{is} \\ 0 & \text{otherwise} \end{cases}$ 
end for
for all  $i \notin \mathcal{I}_k$  do
   $f_{i,r}^{k+1} = \begin{cases} 1 & \text{if } h_{i,r} > 0 \\ 0 & \text{otherwise} \end{cases}$ 
end for

```

**Output:** Next iterate  $F^{k+1} = (f_1^{k+1}, \dots, f_R^{k+1})$

---

In case of a tie we break it randomly. On the other hand, if  $i \notin \mathcal{I}_k$  we have:

$$f_{i,r}^{k+1} = \begin{cases} 1 & \text{if } \nabla e(f_r^k)_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

So each vertex  $i \in \mathcal{I}_k$  is assigned to exactly one and only one cluster, whereas a vertex  $i \notin \mathcal{I}_k$  can be assigned to multiple clusters or no cluster at all. The gradient of  $e(f)$  is given by the formula:

$$\nabla e(f) = \log \left( \frac{M_\alpha^{-1} f}{\langle f, \mathbf{1}_V \rangle} \right) + (M_\alpha^{-1})^T g - \mathbf{1}_V \quad \text{where} \quad g_i := \frac{f_i}{(M_\alpha^{-1} f)_i}. \quad (11)$$

So the Linear Program (1)-(3) is solved by Algorithm 1 above. The first loop computes

$$h_r = \nabla e(f_r) \quad \text{for } r = 1, \dots, R$$

according to formula (11), and the two following loops perform the gradient thresholding according to formula (9) and (10). The Randomized SLP algorithm for PCut presented in the main body of the paper is obtained by simply adding an outer loop over the variable  $k$  to the above algorithm.

## 2 Algorithmic Details: Algebraic Multigrid

This section details our approach for approximately solving the linear systems

$$M_\alpha x = b \quad \text{and} \quad M_\alpha^{-1} x = b$$

required by our algorithm. Let  $W$  denote an  $n \times n$  matrix with non-negative entries and  $P = D^{-1}W$  for  $D$  the diagonal matrix of vertex degrees. Note that  $P$  is row stochastic, so that  $P\mathbf{1} = \mathbf{1}$  holds. We create a hierarchy  $P_{1 \leq \ell \leq L}^\ell$  of row-stochastic matrices, together with inter-grid transfer operators  $(I^\ell, R^\ell)$  in the following way.

Starting from  $P = D^{-1}W$ , we first create a set of  $n_c \approx n/2$  coarse-level vertices through a simple pair-wise merging of vertices. At the most abstract level, this involves the following process:

- (i) Choose some permutation  $\pi : \{1, \dots, n\} \mapsto \{1, \dots, n\}$  of the vertices. Initialize empty lists  $V_{\text{parent}}$  and  $V_{\text{child}}$  of parent and child vertices. Initialize  $V_{\text{orphan}} = \{1, \dots, n\}$  as a list of un-paired vertices.
- (ii) Visit each vertex according to the ordering given by  $\pi$ , and if  $\pi(i) \in V_{\text{orphan}}$  (i.e.  $\pi(i)$  is currently un-paired) then

- (a) If  $\pi(i)$  has no un-paired column neighbor, then leave  $\pi(i)$  in the  $V_{\text{orphan}}$  list and continue to the next vertex.
- (b) Otherwise, select an un-paired column neighbor  $j_* \in \{j : P_{j,\pi(i)} > 0, j \in V_{\text{orphan}}\}$  of  $\pi(i)$  according to some criterion

$$j_* = \operatorname{argmax}_{\{j: P_{j,\pi(i)} > 0, j \in V_{\text{orphan}}\}} F(\pi(i), j).$$

Push back  $\pi(i)$  onto  $V_{\text{parent}}$  and  $j_*$  onto  $V_{\text{child}}$ , then remove both  $j_*$  and  $\pi(i)$  from the  $V_{\text{orphan}}$  list.

We generally choose

$$F(i, j) = f(P_{ij}, P_{ji}, d_i, d_j)$$

as some function of the weights and vertex degrees, such as  $F(i, j) = P_{ij} + P_{ji}$  for instance.

At the end of the process, we have a list of (parent, child) pairs and the remaining list  $V_{\text{orphan}}$  of unmarked vertices. Let  $n_p$  denote the number of parents and  $n_o$  the number of orphans. Set  $n_c = n_p + n_o$  as the total number of coarse-level vertices. We then use the result of this merging process to create an  $n_c \times n$  restriction matrix  $R$  and an  $n \times n_c$  interpolation matrix  $I$  as follows. Each coarse level vertex  $1 \leq i \leq n_c$  (listed in some arbitrary order) corresponds to either a (parent, child) pair  $(j_1, j_2)$  with  $1 \leq j_1, j_2 \leq n$  of fine-level vertices or a singleton  $1 \leq j_1 \leq n$  orphan vertex. We define the  $n \times n_c$  prolongation or interpolation matrix  $I$  by a simple copy procedure. If a coarse level vertex  $i \leftrightarrow (j_1, j_2)$  corresponds to a pair of fine-level vertices, we set

$$I_{j_1, i} = 1 \quad \text{and} \quad I_{j_2, i} = 1$$

and all other entries  $I_{ki}$  of the  $i^{\text{th}}$  column to zero. If  $i \leftrightarrow j_1$  corresponds to an orphan vertex we set  $I_{j_1, i} = 1$  and all other entries  $I_{ki}$  of the  $i^{\text{th}}$  column to zero. Thus each column of  $I$  contains either one or two non-zero entries, while each row of  $I$  contains a single non-zero entry. Exact interpolation of constants  $I\mathbf{1} = \mathbf{1}$  therefore holds. We define the restriction

$$R = (\operatorname{diag}(I^T \mathbf{1}))^{-1} I^T$$

as the transpose of prolongation, followed by a row-normalization. Thus  $R$  either averages values or copies values of fine level vertices, depending on whether a coarse level vertex corresponds to a pair or a singleton. Finally we use

$$P_c := RPI$$

for the coarse level weights. Note  $P_c$  is entri-wise positive, and moreover

$$P_c \mathbf{1} = RPI\mathbf{1} = RP\mathbf{1} = R\mathbf{1} = \mathbf{1}$$

so row-stochasticity is preserved. A simple calculation shows that

$$RI = \operatorname{Id},$$

and so restriction provides a left-inverse for interpolation. Given the output  $P^2 = P_c$  of the first coarsening, we then iteratively apply this coarsening strategy to obtain a sequence of coarsened matrices  $P^\ell$  and an  $L$ -level hierarchy  $\{P^\ell\}_{1 \leq \ell \leq L}$  of row-stochastic weight matrices. We adopt the convention that  $P^1 = P$  is the original weight matrix, so  $P^L$  corresponds to the coarsest level weights. We terminate the procedure when  $P^L$  contains no more than 500 rows. We also have a collection  $\{I^\ell, R^\ell\}_{2 \leq \ell \leq L}$  of inter-level interpolation and restriction operators, so  $P^\ell = R^\ell P^{\ell-1} I^\ell$  according to our convention.

We then use this hierarchy to approximate the solution of the linear systems

$$(\operatorname{Id} - \alpha P)x = b,$$

where  $0 < \alpha < 1$  the random-walk parameter. We accomplish this by applying a sequence of “half V-cycles.” A single “half V-cycle” consists of the following steps.

- (i) Compute the current residual  $r = b - x + \alpha Px$ .
- (ii) Restrict the current residual to the coarsest level  $r^L = R^L R^{L-1} \dots R^2 r$ .
- (iii) Solve  $(\operatorname{Id} - \alpha P^L)e^L = r^L$  exactly for the error on the coarsest level.
- (iv) For  $\ell = L - 1$  to  $\ell = 1$ ,

- (a) Interpolate the error  $e^\ell = I^{\ell+1}e^{\ell+1}$
- (b) Perform  $k_\ell$  iterations of Gauss-Seidel on the system  $(\text{Id} - \alpha P)e = r^\ell$  with  $e^\ell$  as initialization, and update  $e^\ell$  to the result.
- (v) Correct  $x \leftarrow x + e^1$

In practice we find one iteration of this process is enough for our purposes, in that this is generally sufficient to propagate information across the full graph. We also take  $k_\ell = 1$  for all levels of the hierarchy in our experiments.

To solve the transposed system

$$(\text{Id} - \alpha P^T)x = b,$$

we note that  $P = D^{-1}W$  for  $W = W^T$  a symmetric weight matrix and  $D$  the diagonal matrix of vertex degrees. By a simple change of variables we have

$$(\text{Id} - \alpha P^T)^{-1} = D(\text{Id} - \alpha P)^{-1}D^{-1}.$$

We therefore slightly modify the half V-cycle strategy. We use  $r^1 := D^{-1}(\text{Id} - \alpha D^{-1}W)x$  as the initial residual, apply the half V-cycle above for the system  $(\text{Id} - \alpha P)e^1 = r^1$ , then apply  $x \leftarrow x + De^1$  to update the result.

### 3 Algorithmic Details: Dataset Construction

- 20NEWS (unweighted similarity matrix): The word count from the raw documents was computed using the Rainbow library [1] with a default list of stop words. Words appearing less than 20 times were also removed. The similarity matrix was then obtained by 5 nearest neighbors using cosine similarity between tf-idf features. Source: <http://www.cs.cmu.edu/~mccallum/bow/rainbow/>
- RCV1 (weighted similarity matrix): This dataset was obtained in preprocessed format from <http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html> with the tf-idf features were already computed. We then simply used cosine similarity and 5-NN.
- WEBKB4 (unweighted similarity matrix): The word count from the raw documents was done with the Rainbow library [1]. A list of stop word was removed. Words appearing less than 5 times were removed. The similarity matrix was then obtained by 5 nearest neighbors using cosine similarity between tf-idf features. Source: <http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/>
- CITESEER (weighted similarity matrix): This dataset was obtained in preprocessed format from <http://lincs.cs.umd.edu/projects/projects/lbc/index.html> where each publication in the dataset is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. We then simply used cosine similarity and 5-NN.
- MNIST, PENDIGITS, OPTDIGITS (unweighted similarity matrix): The similarity matrices were obtained from [2], where the authors first extracted scattering features using [3] for images before calculating the 10-NN graph. Source: <http://users.ics.aalto.fi/rozyang/nmfr/index.shtml>
- USPS (weighted similarity matrix): We computed a 10-NN graph using standard Euclidean distance between the raw images. Each edge in the 10-NN graph was given the weight

$$w_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

where each  $\mathbf{x}_i$  denotes a vector containing the raw pixel data. The parameter  $\sigma$  was chosen as the mean distance between each vertex and its 10<sup>th</sup> nearest neighbor. Source: <http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html>

## 4 Theoretical Details: Equivalence of (4) and (5) with Pcut and Ncut

The fact that maximizing the geometric average  $(\prod_r \prod_{i \in A_r} \mathbf{pr}_{A_r}(v_i))^{1/n}$  is equivalent to minimizing  $\mathbf{Pcut}(\mathcal{P})$  simply comes from the equality:

$$\mathbf{Pcut}(\mathcal{P}) = \frac{(\prod_{i \in V} \mathbf{pr}_V(v_i))^{1/n}}{(\prod_r \prod_{i \in A_r} \mathbf{pr}_{A_r}(v_i))^{1/n}}. \quad (12)$$

This equality can easily be verified by using the fact that  $\mathbf{pr}_A(v_i) = \frac{1}{|A|} \sum_{j \in A} \omega_{ij}$ . To show that maximizing the arithmetic average  $\frac{1}{n} \sum_r \sum_{i \in A_r} \mathbf{pr}_{A_r}(v_i)$  is equivalent to minimizing  $\mathbf{Ncut}(\mathcal{P})$ , note that:

$$\begin{aligned} \sum_{r=1}^R \sum_{i \in A_r} \mathbf{pr}_{A_r}(v_i) &= \sum_{r=1}^R \left( 1 - \sum_{i \in A_r^c} \mathbf{pr}_{A_r}(v_i) \right) = R - \sum_{r=1}^R \frac{\sum_{i \in A_r^c} \sum_{j \in A_r} \omega_{ij}}{|A_r|} \\ &= R - \sum_{r=1}^R \frac{\sum_{i \in A_r^c} \sum_{j \in A_r} \omega_{ij}}{\sum_{i \in V} \sum_{j \in A_r} \omega_{ij}} \end{aligned} \quad (13)$$

## 5 Theoretical Details: Proof of Theorem 1

In this section we prove the inequality:

$$e^{-H(\mathcal{P})} \leq \mathbf{Pcut}(\mathcal{P}) \leq 1. \quad (14)$$

The upper bound can be directly read from the definition of the Product Cut. Using (12), we see that the upperbound is equivalent to

$$\left( \prod_{i \in V} \mathbf{pr}_V(v_i) \right)^{1/n} \leq \left( \prod_r \prod_{i \in A_r} \mathbf{pr}_{A_r}(v_i) \right)^{1/n} \quad (15)$$

Taking the logarithm of both sides, and using the fact that

$$e(\mathbf{1}_{A_r}) = \sum_{i \in A_r} \log \frac{(M_\alpha^{-1} \mathbf{1}_{A_r})_i}{|A_r|} = \sum_{i \in A_r} \log (\mathbf{pr}_{A_r}(v_i))$$

where  $e(f)$  is the energy defined in section 3 of the paper, we see that (15) is equivalent to

$$e(\mathbf{1}_V) \leq \sum_r e(\mathbf{1}_{A_r})$$

Since  $e(f)$  is 1-homogeneous and convex (see Theorem 3),  $e(f)$  is subadditive and therefore

$$e(\mathbf{1}_V) = e\left(\sum_r \mathbf{1}_{A_r}\right) \leq \sum_r e(\mathbf{1}_{A_r})$$

## 6 Theoretical Details: Proof of Theorem 2

The unperturbed graph  $\mathcal{G}_n = (V_n, W_n)$  was constructed in a way so that the personalized page-rank vectors  $\mathbf{pr}_{A_n}$  and  $\mathbf{pr}_{B_n}$  can be explicitly computed. The formula for  $\mathbf{pr}_{A_n}$  is provided in the following lemma. The formula for  $\mathbf{pr}_{B_n}$  is obtained by exchanging the role of  $A_n$  and  $B_n$ .

**Lemma 1** *The personalized page-rank vector  $\mathbf{pr}_{A_n}$  on the unperturbed graph  $\mathcal{G}_n = (V_n, W_n)$  is given by the formula:*

$$\mathbf{pr}_{A_n} = x \frac{\mathbf{1}_{A_n}}{|A_n|} + y \frac{\mathbf{1}_{B_n}}{|B_n|} \quad \text{where} \quad x = \frac{1 - \alpha + \alpha\mu}{1 - \alpha + 2\alpha\mu} \quad \text{and} \quad y = 1 - x.$$

**Proof.** The solution  $u = M_\alpha^{-1} \mathbf{1}_A$  satisfies

$$u = \alpha W D^{-1} u + (1 - \alpha) \mathbf{1}_A \quad (16)$$

where we use have dropped the subscript on  $W_n$  for simplicity of notation. Since each vertex in  $\mathcal{G}_n$  has degree  $k$  and is connected to  $\mu k$  vertices in the opposite cluster, we have that:

$$W D^{-1} \mathbf{1}_A = (1 - \mu) \mathbf{1}_A + \mu \mathbf{1}_B \quad (17)$$

$$W D^{-1} \mathbf{1}_B = \mu \mathbf{1}_A + (1 - \mu) \mathbf{1}_B \quad (18)$$

Let us look for a solution  $u$  of equation (16) that has the form  $u = x \mathbf{1}_A + y \mathbf{1}_B$ . For such a  $u$  we have

$$W D^{-1} u = (x(1 - \mu) + y\mu) \mathbf{1}_A + (x\mu + y(1 - \mu)) \mathbf{1}_B$$

and equation (16) implies:

$$\begin{cases} x = \alpha(x(1 - \mu) + y\mu) + (1 - \alpha) \\ y = \alpha(x\mu + y(1 - \mu)) \end{cases} \quad \text{or} \quad \begin{cases} x + y = 1 \\ \alpha\mu x + (\alpha - \alpha\mu - 1)y = 0 \end{cases}$$

The solution of this system is:

$$x = \frac{1 - \alpha + \alpha\mu}{1 - \alpha + 2\alpha\mu}$$

The formula for  $\mathbf{pr}_{A_n} = M_\alpha^{-1} \mathbf{1}_{A_n} / |A_n|$  is obtained by scaling  $u$  by a factor  $1/|A_n|$  and by noting that  $A_n = B_n$ .  $\square$

We will also need an explicit formula for  $\mathbf{pr}_{V_n}$  on the unperturbed graph, but this one is trivial:

**Lemma 2**  $\mathbf{pr}_{V_n} = \frac{\mathbf{1}_{V_n}}{|V_n|}$

We next get an estimate for the personalized page-rank vector  $\mathbf{pr}_C^0$  on the perturbed graph  $\mathcal{G}_n^0$ :

**Lemma 3**  $\mathbf{pr}_C^0(v_i) \geq \frac{1}{|C|} \frac{1 - \alpha}{1 - \alpha + \alpha\mu_0}$  for all  $v_i \in C$ .

**Proof.** Consider the two following Markov processes on the graph  $\mathcal{G}_n^0$ :

- Process 1: At each step, the random walker has a probability  $1 - \alpha$  to be teleported to the set  $C$ , and a probability  $\alpha$  to move to a neighboring vertex via a step of regular graph random walk.
- Process 2: The directed version of process 1 where edges connecting  $C$  to  $C^c$  only exit  $C$ . In other words, when performing a step of regular random walk, the random walker can use the edges connecting  $C$  to  $C^c$  only to exit  $C$ . Once the random walker is outside of  $C$ , the only way he can come back to  $C$  is by teleportation.

Process 1 is the process associated with personalized page-rank. The stationary distribution of process 1 is the vector  $\mathbf{pr}_C^0$ . Let  $\mathbf{pr}_C^{0,\text{mod}}$  be the stationary distribution of the process 2. Since the only difference between process 1 and process 2 is that when the random walker is outside of  $C$ , he is less likely to comeback in  $C$ , it is clear that

$$\mathbf{pr}_C^0(v_i) \geq \mathbf{pr}_C^{0,\text{mod}}(v_i) \quad \text{for all } v_i \in C.$$

Since all the vertices in  $C$  are equivalent (due to the graph construction), Process 2 reduces to a two states process, where the first state is “in C” and the second state is “out C”. The (column stochastic) transition matrix is of this two states process is:

$$N = \begin{bmatrix} \alpha(1 - \mu_0) + (1 - \alpha) & 1 - \alpha \\ \alpha\mu_0 & \alpha \end{bmatrix} = \begin{bmatrix} 1 - \alpha\mu_0 & 1 - \alpha \\ \alpha\mu_0 & \alpha \end{bmatrix}$$

and the stationary distribution (that is, the solution of  $Nu = u$ ), is

$$u = \begin{bmatrix} \frac{1 - \alpha}{1 - \alpha + \alpha\mu_0} \\ \frac{\alpha\mu_0}{1 - \alpha + \alpha\mu_0} \end{bmatrix}$$

Again, since all the vertices in  $C$  are equivalent, we obtain

$$\mathbf{pr}_C^{0,\text{mod}}(v_i) = \frac{1}{|C|} \frac{1 - \alpha}{1 - \alpha + \alpha\mu_0} \quad \text{for all } v_i \in C.$$

which concludes the proof.  $\square$

The following lemma is quite technical. Its proof can be found in the next section.

**Lemma 4 (First Technical Lemma)**

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in V_n} \mathbf{pr}_{V_n}^0(v_i) - \sum_{v_i \in V_n} \mathbf{pr}_{V_n}(v_i) \right| = 0 \quad (19)$$

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in A_n} \mathbf{pr}_{A_n}^0(v_i) - \sum_{v_i \in A_n} \mathbf{pr}_{A_n}(v_i) \right| = 0 \quad (20)$$

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in B_n} \mathbf{pr}_{B_n \cup C}^0(v_i) - \sum_{v_i \in B_n} \mathbf{pr}_{B_n}(v_i) \right| = 0 \quad (21)$$

$$\lim_{n \rightarrow \infty} \sum_{v_i \in C} \mathbf{pr}_{B_n \cup C}^0(v_i) = 0 \quad (22)$$

We are now ready to prove the normalized cut is not stable, which is the first part of Theorem 2:

**Proposition 1** *Suppose that  $\mu, \mu_0, k, k_0, n_0$  are fixed. Then*

$$\mu_0 < 2\mu \quad \Rightarrow \quad \mathbf{Ncut}(\mathcal{P}_n^{0,\text{good}}) > \mathbf{Ncut}(\mathcal{P}_n^{0,\text{bad}}) \quad \text{for } n \text{ large enough.} \quad (23)$$

**Proof.** Define:

$$E_n^0(\mathcal{P}_n^{0,\text{good}}) = E_n^0(A_n, B_n \cup C) = \sum_{v_i \in A_n} \mathbf{pr}_{A_n}^0(v_i) + \sum_{v_i \in B_n \cup C} \mathbf{pr}_{B_n \cup C}^0(v_i) \quad (24)$$

$$E_n^0(\mathcal{P}_n^{0,\text{bad}}) = E_n^0(A_n \cup B_n, C) = \sum_{v_i \in A_n \cup B_n} \mathbf{pr}_{A_n \cup B_n}^0(v_i) + \sum_{v_i \in C} \mathbf{pr}_C^0(v_i) \quad (25)$$

Since minimizing the Normalized Cut is equivalent to maximizing the arithmetic average of the personalized page-rank vectors, we need to show that  $\mu_0 < 2\mu$  implies that  $E_n^0(\mathcal{P}_n^{0,\text{good}}) > E_n^0(\mathcal{P}_n^{0,\text{bad}})$  for  $n$  large enough.

Since  $\sum_{v_i \in V_n} \mathbf{pr}_{V_n}(v_i) = 1$ , convergence (40) from lemma (4) simply states that

$$\lim_{n \rightarrow \infty} \sum_{v_i \in A_n \cup B_n} \mathbf{pr}_{A_n \cup B_n}^0(v_i) = 1.$$

Combining this with the lower bound from lemma (3) on  $\mathbf{pr}_C^0$  we get:

$$E_n^0(\mathcal{P}_n^{0,\text{bad}}) \geq 1 - \epsilon(n) + \frac{1 - \alpha}{1 - \alpha + \alpha\mu_0} \quad (26)$$

for some function  $\epsilon(n)$  satisfying  $\lim_{n \rightarrow \infty} \epsilon(n) = 0$ . Consider now the energy of the good partition:

$$E_n^0(\mathcal{P}_n^{0,\text{good}}) = \sum_{v_i \in A_n} \mathbf{pr}_{A_n}^0(v_i) + \sum_{v_i \in B_n} \mathbf{pr}_{B_n \cup C}^0(v_i) + \sum_{v_i \in C} \mathbf{pr}_{B_n \cup C}^0(v_i) \quad (27)$$

Using lemma 1 we see that, on the unperturbed graph, we have:

$$\sum_{v_i \in A_n} \mathbf{pr}_{A_n}(v_i) = \sum_{v_i \in B_n} \mathbf{pr}_{B_n}(v_i) = \frac{1 - \alpha + \alpha\mu}{1 - \alpha + 2\alpha\mu}$$

Combining this with convergences (41), (42) and (43) from lemma (4) leads to:

$$\lim_{n \rightarrow \infty} \sum_{v_i \in A_n} \mathbf{pr}_{A_n}^0(v_i) = \lim_{n \rightarrow \infty} \sum_{v_i \in A_n} \mathbf{pr}_{B_n \cup C}^0(v_i) = \sum_{v_i \in A_n} \mathbf{pr}_{A_n}(v_i) = \frac{1 - \alpha + \alpha\mu}{1 - \alpha + 2\alpha\mu}$$

together with the fact that  $\lim_{n \rightarrow \infty} \sum_{v_i \in C} \mathbf{pr}_{B_n \cup C}^0(v_i) = 0$  we obtain that

$$\lim_{n \rightarrow \infty} E^0(\mathcal{P}_n^{0, \text{good}}) = 2 \frac{1 - \alpha + \mu\alpha}{1 - \alpha + 2\mu\alpha} = 1 + \frac{1 - \alpha}{1 - \alpha + 2\mu\alpha} \quad (28)$$

Comparing (26) and (28) and noticing that  $\mu_0 < 2\mu$  implies that  $\frac{1-\alpha}{1-\alpha+\mu_0\alpha} > \frac{1-\alpha}{1-\alpha+2\mu\alpha}$  conclude the proof.  $\square$

From lemma 1 and 2 we know that

$$2n \mathbf{pr}_{A_n}(v_i) = 2x \quad \text{for all } v_i \in A_n \quad (29)$$

$$2n \mathbf{pr}_{B_n}(v_i) = 2x \quad \text{for all } v_i \in B_n \quad (30)$$

$$2n \mathbf{pr}_{V_n}(v_i) = 1 \quad \text{for all } v_i \in V_n. \quad (31)$$

where  $x$  is the quantity defined in lemma 1. The following technical lemma, whose proof can be found in the next section, is require to prove the Product Cut is stable.

**Lemma 5 (Second Technical Lemma)**

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in A_n} 2n \mathbf{pr}_{A_n}^0(v_i) \prod_{i \in B_n} 2n \mathbf{pr}_{B_n \cup C}^0(v_i) \right)^{1/(2n)} = 2x \quad (32)$$

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in V_n} 2n \mathbf{pr}_{V_n \cup C}^0(v_i) \right)^{1/(2n)} = 1 \quad (33)$$

We are now ready to prove the Product Cut is stable, which is the second part of Theorem 2:

**Proposition 2** *Suppose that  $\mu, \mu_0, k, k_0, n_0$  are fixed. Then*

$$\mathbf{Pcut}(\mathcal{P}_n^{0, \text{good}}) < \mathbf{Pcut}(\mathcal{P}_n^{0, \text{bad}}) \quad \text{for } n \text{ large enough.} \quad (34)$$

The sequence of partitions  $\mathcal{P}_n^{0, \text{bad}}$  becomes arbitrarily ill-balanced, which from (14) implies the following limit on the perturbed graph  $\mathcal{G}_n^0$ :

$$\lim_{n \rightarrow \infty} \mathbf{Pcut}_{\mathcal{G}_n^0}(\mathcal{P}_n^{0, \text{bad}}) = 1. \quad (35)$$

From equation (12), lemma 1 and lemma 2, we have that

$$\mathbf{Pcut}_{\mathcal{G}_n}(A_n, B_n) = \frac{1}{2x} < 1 \quad (36)$$

where  $x$  is defined in lemma (1) In order to conclude we will show that

$$\lim_{n \rightarrow \infty} \mathbf{Pcut}_{\mathcal{G}_n^0}(A_n, B_n \cup C) = \frac{1}{2x}. \quad (37)$$

Indeed, combined with (35), since  $1/(2x) < 1$ , the above limit shows that the Product Cut of  $\mathcal{P}_n^{0, \text{good}} = (A_n, B_n \cup C)$  becomes eventually smaller than the Product Cut of  $\mathcal{P}_n^{0, \text{bad}}$ .



We now prove (37). Using equality (12), and noting that the perturbed graph  $\mathcal{G}_n^0$  has  $2n + n_0$  vertices, we have:

$$\begin{aligned}
\mathbf{Pcut}_{\mathcal{G}_n^0}(A_n, B_n \cup C) &= \frac{(\prod_{i \in V_n \cup C} \mathbf{pr}_{V_n \cup C}^0(v_i))^{1/(2n+n_0)}}{(\prod_{i \in A_n} \mathbf{pr}_{A_n}^0(v_i) \prod_{i \in B_n \cup C} \mathbf{pr}_{B_n \cup C}^0(v_i))^{1/(2n+n_0)}} \\
&= \frac{(\prod_{i \in V_n \cup C} 2n \mathbf{pr}_{V_n \cup C}^0(v_i))^{1/(2n+n_0)}}{(\prod_{i \in A_n} 2n \mathbf{pr}_{A_n}^0(v_i) \prod_{i \in B_n \cup C} 2n \mathbf{pr}_{B_n \cup C}^0(v_i))^{1/(2n+n_0)}} \\
&= \frac{\left[ (\prod_{i \in V_n} 2n \mathbf{pr}_{V_n \cup C}^0(v_i))^{1/(2n)} \right]^{2n/(2n+n_0)} \left[ \prod_{i \in C} 2n \mathbf{pr}_{V_n \cup C}^0(v_i) \right]^{1/(2n+n_0)}}{\left[ (\prod_{i \in A_n} 2n \mathbf{pr}_{A_n}^0(v_i) \prod_{i \in B_n} 2n \mathbf{pr}_{B_n \cup C}^0(v_i))^{1/(2n)} \right]^{2n/(2n+n_0)} \left[ \prod_{i \in C} 2n \mathbf{pr}_{B_n \cup C}^0(v_i) \right]^{1/(2n+n_0)}} \\
&= \frac{\text{term1} \times \text{term2}}{\text{term3} \times \text{term4}}
\end{aligned}$$

According to lemma 5 term 1 converges to 1 and term 3 converges to  $2x$ . We now show that terms 2 and 4 both converges to 1, which will conclude the proof. Let's prove it from term 4. First note that

$$\mathbf{pr}_{B_n \cup C}^0(v_i) \geq (1 - \alpha)/|B_n \cup C| \quad \text{for all vertex in } B_n \cup C \quad (38)$$

This is a simple consequence of the fact that the vector  $u = \mathbf{pr}_{B_n \cup C}^0 = M_\alpha^{-1} \frac{\mathbf{1}_{B_n \cup C}}{|B_n \cup C|}$  satisfies the equation

$$u = \alpha W D^{-1} u + (1 - \alpha) \frac{\mathbf{1}_{B_n \cup C}}{|B_n \cup C|}$$

Using (38) we get

$$\begin{aligned}
0 \geq \log(\text{term 4}) &= \frac{1}{2n + n_0} \sum_{i \in C} \log(2n \mathbf{pr}_{B_n \cup C}^0(v_i)) \\
&\geq \frac{1}{2n + n_0} |C| \log\left(2n \frac{1 - \alpha}{|B_n \cup C|}\right) = \frac{n_0}{2n + n_0} \log\left(\frac{2n}{n + n_0} (1 - \alpha)\right) \rightarrow 0
\end{aligned}$$

Term 2 can be handled similarly.  $\square$

## 7 Theoretical Details: Proof Lemmas 4 and 5

We begin two intermediate lemmata that, while elementary, will prove useful in proving lemmas 4 and 5. Let  $P$  denote an  $n \times n$  matrix with non-negative entries. We say  $P$  is *column sub-stochastic* if

$$P_{ij} \geq 0 \quad \text{and} \quad \max_{1 \leq j \leq n} \sum_{i=1}^n P_{ij} \leq 1,$$

or in other words if the maximal column sum of  $P$  remains bounded by unity. After recalling the definition of the  $\ell^1 \rightarrow \ell^1$  operator norm  $\|P\|_1$  of a matrix

$$\|P\|_1 := \max_{\{u \in \mathbb{R}^n, \|u\|_{\ell^1}=1\}} \|Pu\|_{\ell^1} = \max_{1 \leq j \leq n} \sum_{i=1}^n |P_{ij}|,$$

we note that a column sub-stochastic matrix  $P$  has norm  $\|P\|_1$  at most one. Similarly, we say  $P$  is *row sub-stochastic* if

$$P_{ij} \geq 0 \quad \text{and} \quad \max_{1 \leq i \leq n} \sum_{j=1}^n P_{ij} \leq 1,$$

and note analogously that any row sub-stochastic matrix  $P$  has  $\ell^\infty$  operator norm  $\|P\|_\infty$  at most one.

**Lemma 6** Let  $P$  denote a column (row) sub-stochastic matrix. Then

$$M_{\alpha,P}^{-1} := (1 - \alpha)(I - \alpha P)^{-1}$$

exists and is also column (row) sub-stochastic. Moreover, for any  $n \times n$  matrices  $P_1$  and  $P_2$ , the identity

$$M_{\alpha,P_2}^{-1} = M_{\alpha,P_1}^{-1} + \frac{\alpha}{1 - \alpha} M_{\alpha,P_2}^{-1} (P_2 - P_1) M_{\alpha,P_1}^{-1},$$

holds whenever both inverses exist. As a consequence, for any  $u \in \mathbb{R}^n$  and any  $1 \leq p \leq \infty$  the corresponding estimates

$$\|M_{\alpha,P_2}^{-1} u - M_{\alpha,P_1}^{-1} u\|_{\ell^p} \leq \frac{\alpha}{1 - \alpha} \|M_{\alpha,P_2}^{-1} (P_2 - P_1) M_{\alpha,P_1}^{-1} u\|_{\ell^p}$$

hold.

**proof** The first assertion is well-known and standard. It follows, for instance, by appealing to the convergent Neumann series representation

$$M_{\alpha,P}^{-1} = (1 - \alpha) \left( I + \sum_{k=1}^{\infty} \alpha^k P^k \right)$$

for the inverse. The second statement follows from observing that

$$(I - \alpha P_2)^{-1} = [(I - \alpha P_1)^{-1} + \alpha(I - \alpha P_2)^{-1} (P_2 - P_1) (I - \alpha P_1)^{-1}],$$

then simply applying the first part of the lemma.  $\square$

We shall use this lemma to estimate the difference between personalized page-rank vectors computed on the original and perturbed graphs. Let  $V_n := \{1, \dots, n\}$  denote the original vertex set and  $C := \{n+1, \dots, n+n_0\}$  the perturbation set of vertices. Take  $A \subset \{1, \dots, n, n+1, \dots, n+n_0\}$  arbitrary and decompose its indicator  $\mathbf{1}_A$  as

$$\mathbf{1}_A = \begin{pmatrix} \mathbf{1}_{A \cap V_n} \\ \mathbf{1}_{A \cap C} \end{pmatrix}$$

and similarly let

$$\begin{aligned} \|u\|_{\ell^1(V_n)} &:= \sum_{i=1}^n |u_i| & \|u\|_{\ell^1(C)} &:= \sum_{i=n+1}^{n+n_0} |u_i| \\ \|u\|_{\ell^\infty(V_n)} &:= \max_{1 \leq i \leq n} |u_i| & \|u\|_{\ell^\infty(C)} &:= \max_{n+1 \leq i \leq n+n_0} |u_i| \end{aligned}$$

denote the corresponding decompositions of vector norms. We shall then use

$$\mathbf{pr}_A := (1 - \alpha)(I - W D^{-1})^{-1} \begin{pmatrix} \mathbf{1}_{A \cap V_n} \\ \mathbf{1}_{A \cap C} \end{pmatrix} \quad \text{and} \quad \mathbf{pr}_A^0 := M_{\alpha,P_2}^{-1} \begin{pmatrix} \mathbf{1}_A \end{pmatrix}$$

to denote the personalized page-rank vectors of  $A$  induced by the original and perturbed graphs, respectively.

To simplify the analysis, let us consider the original  $n \times n$  symmetric weight matrix  $W$  as embedded in the  $(n + n_0) \times (n + n_0)$  matrix  $W^0$  of the perturbed graph, where we order the vertices so that

$$W^0 = \begin{bmatrix} W & 0_{n,n_0} \\ 0_{n_0,n} & 0_{n_0,n_0} \end{bmatrix} + \begin{bmatrix} 0_{n,n} & \tilde{W}_{n,n_0} \\ \tilde{W}_{n_0,n} & \tilde{W}_{n_0,n_0} \end{bmatrix}. \quad (39)$$

The sub-matrix  $\tilde{W}_{n,n_0} = W_{n_0,n}^T$  thus encodes any additional edges between vertices in  $\{1, \dots, n\}$  and the  $n_0$  newly added vertices, while the sub-matrix  $\tilde{W}_{n_0,n_0}$  describes the connectivity relation between the added vertices themselves. We may partition the corresponding degree matrices accordingly, so that

$$D^0 = \begin{bmatrix} D & 0_{n,n_0} \\ 0_{n_0,n} & 0_{n_0,n_0} \end{bmatrix} + \begin{bmatrix} \tilde{D} & 0_{n,n_0} \\ 0_{n_0,n} & \tilde{D}_{n_0} \end{bmatrix}.$$

We use  $D = \text{diag}(W\mathbf{1})$  to denote the degree matrix of the original graph,  $\tilde{D} = \text{diag}(\tilde{W}_{n,n_0}\mathbf{1})$  is the degree perturbation of the original  $n$  vertices and  $D_{n_0} = \text{diag}(\tilde{W}_{n_0,n}\mathbf{1} + \tilde{W}_{n_0,n_0}\mathbf{1})$  is simply the degree matrix of the added vertices. We may then define

$$P_2 = W^0(D^0)^{-1}$$

as the random-walk matrix of the perturbed graph and

$$P_1 = \begin{bmatrix} WD^{-1} & 0_{n,n_0} \\ 0_{n_0,n} & 0_{n_0,n_0} \end{bmatrix}$$

as the  $(n + n_0) \times (n + n_0)$  embedding of the original random walk matrix into the larger vertex set. Finally, we let

$$\Delta_{\text{RW}} := W(D + \tilde{D})^{-1} - WD^{-1}$$

denote the perturbation of the original random walk matrix itself. With these conventions and definitions in place, we may prove the second intermediate lemma.

**Lemma 7** *Let  $W$  denote an  $n \times n$  symmetric matrix and  $W^0$  an  $(n + n_0) \times (n + n_0)$  perturbation of the form (39). Define  $\tilde{B} \subset \{1, \dots, n\}$*

$$\tilde{B} := \{i \in \{1, \dots, n\} : \tilde{D}_{ii} \neq 0\}$$

*as those vertices in  $W$  affected by the perturbation. Then for any  $A \subset V_n \cup C$ , the estimate*

$$\|(1-\alpha)\mathbf{1}_{A \cap C} - |A|\mathbf{pr}_A^0\|_{\ell^1(C)} + \| |A \cap V_n| \mathbf{pr}_A - |A|\mathbf{pr}_A^0 \|_{\ell^1(V_n)} \leq \frac{2\alpha|A \cap V_n|}{1-\alpha} \|\mathbf{pr}_A\|_{\ell^1(\tilde{B})} + \alpha|A \cap C|$$

*holds for the difference between induced page-rank vectors.*

**proof** The fact that  $D_{ii} = (D + \tilde{D})_{ii}$  unless  $i \in \tilde{B}$  implies that

$$(\Delta_{\text{RW}}u)(v_i) = - \sum_{j \in \tilde{B}} w_{ij} \left( \frac{\tilde{D}_{jj}}{D_{jj}(D_{jj} + \tilde{D}_{jj})} \right) u_j.$$

for  $u \in \mathbb{R}^n$  arbitrary. In a similar fashion, the fact that  $\tilde{W}_{n_0,n} = \tilde{W}_{n,n_0}^T$  implies

$$(\tilde{W}_{n_0,n}(D + \tilde{D})^{-1}u)(v_i) = \sum_{j \in \tilde{B}} \frac{(\tilde{W}_{n_0,n})_{ij}u_j}{(D_{jj} + \tilde{D}_{jj})}.$$

Now let  $R$  denote the  $(n + n_0) \times |\tilde{B}|$  matrix with entries

$$R_{ij} = \frac{-w_{ij}\tilde{D}_{jj}}{D_{jj}(D_{jj} + \tilde{D}_{jj})} \quad (1 \leq i \leq n) \quad \text{and} \quad R_{ij} = \frac{(\tilde{W}_{n_0,n})_{ij}u_j}{(D_{jj} + \tilde{D}_{jj})} \quad (n+1 \leq i \leq n+n_0),$$

so that

$$\|\Delta_{\text{RW}}u\|_{\ell^1(V_n)} + \|\tilde{W}_{n_0,n}(D + \tilde{D})^{-1}u\|_{\ell^1(C)} = \|Ru_{\tilde{B}}\|_{\ell^1}.$$

As the maximal column sums of  $R$  are bounded by

$$\max_{j \in \tilde{B}} \frac{2\tilde{D}_{jj}}{(D_{jj} + \tilde{D}_{jj})} \leq 2,$$

the estimate  $\|Ru_{\tilde{B}}\|_{\ell^1} \leq 2\|u\|_{\ell^1(\tilde{B})}$  then follows. Now take  $u \in \mathbb{R}^{n_0}$  arbitrary and recall that the matrix

$$Q := \begin{bmatrix} \tilde{W}_{n,n_0} \tilde{D}_{n_0}^{-1} \\ \tilde{W}_{n_0,n_0} \tilde{D}_{n_0}^{-1} \end{bmatrix}$$

is column stochastic by definition. Thus

$$\|\tilde{W}_{n,n_0} \tilde{D}_{n_0}^{-1}u\|_{\ell^1(V_n)} + \|\tilde{W}_{n_0,n_0} \tilde{D}_{n_0}^{-1}u\|_{\ell^1(C)} = \|Qu\|_{\ell^1(V_n \cup C)} \leq \|u\|_{\ell^1(C)}$$

by definition of the  $\ell^1$ -matrix norm. Combining these estimates with lemma 6 and the choices  $u = M_\alpha^{-1}\mathbf{1}_{A \cap V_n}$  and  $u = (1 - \alpha)\mathbf{1}_{A \cap C}$  then shows

$$\|(1-\alpha)\mathbf{1}_{A \cap C} - |A|\mathbf{pr}_A^0\|_{\ell^1(C)} + \| |A \cap V_n| \mathbf{pr}_A - |A|\mathbf{pr}_A^0 \|_{\ell^1(V_n)} \leq \frac{2\alpha|A \cap V_n|}{1-\alpha} \|\mathbf{pr}_A\|_{\ell^1(\tilde{B})} + \alpha|A \cap C|.$$

which is exactly the claimed bound.  $\square$

With this result in place, we may now prove lemmas 4 and 5

**Lemma 8 (First Technical Lemma)**

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in V_n} \mathbf{pr}_{V_n}^0(v_i) - \sum_{v_i \in V_n} \mathbf{pr}_{V_n}(v_i) \right| = 0 \quad (40)$$

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in A_n} \mathbf{pr}_{A_n}^0(v_i) - \sum_{v_i \in A_n} \mathbf{pr}_{A_n}(v_i) \right| = 0 \quad (41)$$

$$\lim_{n \rightarrow \infty} \left| \sum_{v_i \in B_n} \mathbf{pr}_{B_n \cup C}^0(v_i) - \sum_{v_i \in B_n} \mathbf{pr}_{B_n}(v_i) \right| = 0 \quad (42)$$

$$\lim_{n \rightarrow \infty} \sum_{v_i \in C} \mathbf{pr}_{B_n \cup C}^0(v_i) = 0 \quad (43)$$

**proof** Apply lemma 7 with the choice  $A = V_n$  to find

$$\|\mathbf{pr}_{V_n} - \mathbf{pr}_{V_n}^0\|_{\ell^1(V_n)} \leq C(\alpha) \|\mathbf{pr}_{V_n}\|_{\ell^1(\tilde{B})}.$$

But  $\|\mathbf{pr}_{V_n}\|_{\ell^1(\tilde{B})} = |\tilde{B}|/2n = n_0/2n \rightarrow 0$  since  $n_0$  is constant. This proves the first statement. Applying lemma 7 with  $A = A_n$  yields the second statement in exactly the same way. For the third statement, use the choice  $A = B_n \cup C$  to find

$$\|\mathbf{pr}_{B_n} - (1 + n_0/n)\mathbf{pr}_{B_n \cup C}^0\|_{\ell^1(V_n)} \leq C(\alpha) \|\mathbf{pr}_{B_n}\|_{\ell^1(\tilde{B})} + \alpha|C|/n \leq C(\alpha)(n_0/n) \rightarrow 0.$$

By the triangle inequality,  $\|\mathbf{pr}_{B_n} - \mathbf{pr}_{B_n \cup C}^0\|_{\ell^1(V_n)} \leq \|\mathbf{pr}_{B_n} - (1 + n_0/n)\mathbf{pr}_{B_n \cup C}^0\|_{\ell^1(V_n)} + \|\mathbf{pr}_{B_n \cup C}^0\|_{\ell^1(V_n)}(n_0/n) \rightarrow 0$ , which yields the third claim. The fourth and final claim follows from the choice  $A = B_n \cup C$ , the bound

$$\|(1 - \alpha)\mathbf{1}_C - (n + n_0)\mathbf{pr}_{B_n \cup C}^0\|_{\ell^1(C)} \leq C(\alpha)n_0$$

and the triangle inequality as well.  $\square$

**Lemma 9 (Second Technical Lemma)**

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in A_n} 2n \mathbf{pr}_{A_n}^0(v_i) \prod_{i \in B_n} 2n \mathbf{pr}_{B_n \cup C}^0(v_i) \right)^{1/(2n)} = 2x \quad (44)$$

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in V_n} 2n \mathbf{pr}_{V_n \cup C}^0(v_i) \right)^{1/(2n)} = 1 \quad (45)$$

**proof** By following the proof of the first technical lemma, we may conclude that the estimates

$$\begin{aligned} \sum_{i \in A_n} |x - n\mathbf{pr}_{A_n}^0(v_i)| &\leq C \\ \sum_{i \in B_n} |x - (n + n_0)\mathbf{pr}_{B_n \cup C}^0(v_i)| &\leq C \\ \sum_{i \in V_n} |1 - (2n + n_0)\mathbf{pr}_{V_n \cup C}^0(v_i)| &\leq C \end{aligned}$$

hold, where  $C > 0$  denotes a uniform constant that does not depend upon  $n$ . Let  $\epsilon > 0$  be arbitrary and set

$$A_n^\epsilon := \{v_i \in V_n : |x - n\mathbf{pr}_{A_n}^0(v_i)| \leq \epsilon\}.$$

The uniform bound above shows that  $|(A_n^\epsilon)^c| \leq C/\epsilon$ , and so

$$\left( \prod_{i \in A_n} n\mathbf{pr}_{A_n}^0(v_i) \right)^{1/2n} = \left( \prod_{i \in A_n^\epsilon} n\mathbf{pr}_{A_n}^0(v_i) \right)^{1/2n} \left( \prod_{i \in (A_n^\epsilon)^c} n\mathbf{pr}_{A_n}^0(v_i) \right)^{1/2n} := \mathrm{I} \times \mathrm{II}.$$

For the first term we have

$$(x - \epsilon)^{\frac{|A_n^\epsilon|}{2n}} \leq I \leq (x + \epsilon)^{\frac{|A_n^\epsilon|}{2n}}$$

by the definition of  $A_n^\epsilon$ . For II we know that

$$(1 - \alpha)^{\frac{|(A_n^\epsilon)^c|}{2n}} \leq II \leq (C + x)^{\frac{|(A_n^\epsilon)^c|}{2n}}$$

by the uniform upper bound and the fact that  $v_i \in A_n$  implies a lower bound of  $(1 - \alpha)/n$  for  $\text{pr}_{A_n}^0(v_i)$ . As  $|A_n^\epsilon| \geq n - C/\epsilon$  this shows

$$\sqrt{x - \epsilon} \leq \liminf_{n \rightarrow \infty} \left( \prod_{i \in A_n} n \text{pr}_{A_n}^0(v_i) \right)^{1/2n} \leq \limsup_{n \rightarrow \infty} \left( \prod_{i \in A_n} n \text{pr}_{A_n}^0(v_i) \right)^{1/2n} \leq \sqrt{x + \epsilon}.$$

Thus

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in A_n} n \text{pr}_{A_n}^0(v_i) \right)^{1/2n} = \sqrt{x}$$

since  $\epsilon > 0$  was arbitrary. As  $n_0$  is fixed, by starting from the second uniform bound a similar computation shows

$$\lim_{n \rightarrow \infty} \left( \prod_{i \in B_n} n \text{pr}_{B_n \cup C}^0(v_i) \right)^{1/2n} = \sqrt{x},$$

which combines with the previous computation to yield the first claim. Finally, beginning from the third uniform bound and applying the same argument yields the second claim.  $\square$

## 8 Proof of Theorem 3

In this section we prove that the energy

$$e(f) = \left\langle f, \log \frac{M_\alpha^{-1} f}{\langle f, \mathbf{1} \rangle} \right\rangle$$

is convex on  $\mathbb{R}_+^n$ . We will employ the following notational conventions throughout the proof. For each  $1 \leq i \leq n$  we let  $\mathbf{e}_i$  denote the  $i^{\text{th}}$  standard basis vector,  $\mathbf{0}$  denote the zero vector and  $\mathbf{1} = (1, 1, \dots, 1)^T$  denote the vector of all ones. We reserve  $\text{Id}$  for the identity matrix. Given any vector  $f = (f_1, \dots, f_n)^T \in \mathbb{R}^n$ , we use

$$\text{diag}(f) := \begin{bmatrix} f_1 & 0 & 0 & \cdots & 0 \\ 0 & f_2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & f_{n-1} & 0 \\ 0 & \cdots & 0 & 0 & f_n \end{bmatrix}$$

to denote the corresponding diagonal matrix with the entries of  $f$  along the diagonal. For any such  $f \in \mathbb{R}^n$  we trivially note that

$$\text{diag}(f) \mathbf{1} = f$$

by definition. Finally, we recall that

$$\mathbb{R}_+^n := \{f \in \mathbb{R}^n : \forall i, f_i \geq 0\}$$

denotes the set of non-negative vectors.

Let  $Q$  denote a column-stochastic, invertible matrix with strictly positive entries. That is,  $Q^T \mathbf{1} = \mathbf{1}$  and  $Q_{ij} > 0$  for all pairs  $(i, j)$  of indices. Suppose further that  $Q^{-1}$  is an  $M$ -matrix — there exists a matrix  $B$  with non-negative entries  $B_{ij} \geq 0$  and a scalar  $\sigma > \rho(B)$  so that

$$Q^{-1} = \sigma \text{Id} - B,$$

where  $\rho(B)$  denotes the spectral radius. The condition  $Q^T \mathbf{1} = \mathbf{1}$  then implies  $B^T \mathbf{1} = (\sigma - 1) \mathbf{1}$ , and so  $\sigma > 1$  since  $B$  has non-negative entries. We may therefore decompose

$$Q^{-1} = \frac{1}{1 - \alpha} (\text{Id} - \alpha P) \quad \text{for} \quad 0 < \alpha := \frac{\sigma - 1}{\sigma} < 1,$$

with  $P$  some column-stochastic matrix with non-negative entries. Given any such column-stochastic matrix  $P$  with non-negative entries, we have a convergent power-series representation

$$Q = (1 - \alpha) (\text{Id} - \alpha P)^{-1} = (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n P^n$$

for the inverse. Thus  $Q_{ij} > 0$  for all  $(i, j)$  precisely when the graph determined by the non-zero entries of  $P$  is connected. We refer to any column-stochastic matrix  $P$  with this property as the *random-walk matrix* of a connected graph. Given  $Q$  of this form, in what follows we shall also let

$$Q = \begin{bmatrix} | & | & & | \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \\ | & | & & | \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} - & \mathbf{r}_1 & - \\ - & \mathbf{r}_2 & - \\ & \vdots & \\ - & \mathbf{r}_n & - \end{bmatrix},$$

so that the  $\mathbf{q}_i$  denote the columns of  $Q$  while the  $\mathbf{r}_i$  denote its rows. Before proceeding with the proof of convexity, we first pause to collect a few elementary results regarding matrices  $Q$  of this form in the following lemma.

**Lemma 10** *Let  $Q$  denote an invertible, column-stochastic matrix with positive entries. Then*

- (i)  $Q^{-1}$  is an  $M$ -matrix if and only if there exists  $0 < \alpha < 1$  so that

$$Q^{-1} = \frac{1}{1 - \alpha} (\text{Id} - \alpha P),$$

where  $P$  is the random-walk matrix of a connected graph.

- (ii) If  $f \in \mathbb{R}_+^n$  and  $f \neq \mathbf{0}$  then  $\langle Qf, \mathbf{e}_i \rangle > 0$  for all  $1 \leq i \leq n$ .  
(iii) For all  $f \in \mathbb{R}^n$ , the conservation of mass property  $\langle Qf, \mathbf{1} \rangle = \langle f, \mathbf{1} \rangle$  holds.

Given a scalar  $\alpha \in (0, 1)$  and a random-walk matrix  $P$  from a connected graph, we let matrix  $Q = (1 - \alpha)(\text{Id} - \alpha P)^{-1}$  denote the corresponding diffused matrix. For any non-zero  $f \in \mathbb{R}_+^n$  we define the corresponding cluster energy  $e(f)$  as

$$e(f) := \sum_{i=1}^n f_i \log \left( \frac{\langle Qf, \mathbf{e}_i \rangle}{\langle f, \mathbf{1} \rangle} \right) = -\langle f, \mathbf{1} \rangle \log \langle f, \mathbf{1} \rangle + \sum_{i=1}^n f_i \log \langle Qf, \mathbf{e}_i \rangle,$$

and we set  $e(\mathbf{0}) = 0$  otherwise. We wish to show that  $e(f)$  restricted to  $\mathbb{R}_+^n$  defines a convex function over a convex set. We accomplish this by means of the following lemma. We shall eventually reduce showing positive semi-definiteness of the Hessian of  $e(f)$  to a direct appeal to this lemma.

**Lemma 11** *Let  $Q = (1 - \alpha)(\text{Id} - \alpha P)^{-1}$  for  $P$  an arbitrary random-walk matrix. Suppose that  $y \in \text{int}(\mathbb{R}_+^n)$  has strictly positive entries. Define  $z \in \text{int}(\mathbb{R}_+^n)$  by  $z_i := y_i^2$  for each  $1 \leq i \leq n$  and  $D_y := \text{diag}(y)$ . Then the matrix*

$$M(y) := D_y Q^{-T} D_y^{-1} + D_y^{-1} Q^{-1} D_y - D_y^{-1} \text{diag}(Q^{-1} z) D_y^{-1} - \frac{yy^T}{y^T y}$$

is positive semi-definite, and

$$\ker(M(y)) = \text{Span}(y).$$

Moreover, the relation

$$\langle v, M(y)v \rangle = |v|_2^2 + \frac{\alpha}{1 - \alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left( \frac{v_i y_j}{y_i} - v_j \right)^2$$

holds for  $v \in \text{Span}^\perp(y)$ .

**proof** The elementary facts  $D_y^{-1} y = \mathbf{1}$ ,  $Q^{-T} \mathbf{1} = \mathbf{1}$  and  $D_y \mathbf{1} = y$  combine to show

$$\left( D_y Q^{-T} D_y^{-1} - \frac{yy^T}{\|y\|_2^2} \right) y = y - y \left( \frac{y^T y}{y^T y} \right) = 0.$$

In a similar fashion, the elementary facts  $D_y y = z$  and  $\text{diag}(Q^{-1}z)\mathbf{1} = Q^{-1}z$  yield

$$(D_y^{-1}Q^{-1}D_y - D_y^{-1}\text{diag}(Q^{-1}z)D_y^{-1})y = D_y^{-1}Q^{-1}z - D_y^{-1}\text{diag}(Q^{-1}z)\mathbf{1} = 0,$$

so that  $y \in \ker(M(y))$  as claimed. Now suppose  $v \in \mathbb{R}^n$  satisfies  $v \perp y$ , so that

$$\langle v, M(y)v \rangle = 2\langle D_y^{-1}Q^{-1}D_y v, v \rangle - \langle D_y^{-1}\text{diag}(Q^{-1}z)D_y^{-1}v, v \rangle := 2\text{I} - \text{II},$$

and recall that

$$Q^{-1} = \frac{1}{1-\alpha}(\text{Id} - \alpha P)$$

for some  $0 < \alpha < 1$ . An entriwise computation shows that the first term I equals

$$\text{I} = \frac{1}{1-\alpha}|v|_2^2 - \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \frac{y_j}{y_i} v_i v_j,$$

while the second term II must satisfy

$$\begin{aligned} \text{II} &= \frac{1}{1-\alpha}|v|_2^2 - \frac{\alpha}{1-\alpha} \sum_{i=1}^n \left( \frac{v_i}{y_i} \right)^2 (Pz)_i \\ &= \frac{1}{1-\alpha}|v|_2^2 - \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left( \frac{v_i y_j}{y_i} \right)^2. \end{aligned}$$

Combining these relations then shows that

$$\begin{aligned} \langle v, M(y)v \rangle &= \frac{1}{1-\alpha}|v|^2 + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left[ \left( \frac{v_i y_j}{y_i} \right)^2 - 2 \frac{y_j}{y_i} v_i v_j \right] \\ &= \frac{1}{1-\alpha}|v|^2 + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left[ \left( \frac{v_i y_j}{y_i} - v_j \right)^2 - v_j^2 \right] \\ &= \frac{1}{1-\alpha}|v|^2 + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left( \frac{v_i y_j}{y_i} - v_j \right)^2 - \frac{\alpha}{1-\alpha} \sum_{j=1}^n v_j^2, \end{aligned}$$

where the last equality follows from the column-stochasticity of  $P$  after reversing the order of summation. For any  $v \perp y$  it therefore follows that

$$\langle v, M(y)v \rangle = |v|_2^2 + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left( \frac{v_i y_j}{y_i} - v_j \right)^2 > 0$$

unless  $v = 0$ , which simultaneously yields positive semi-definiteness of  $M(y)$  and the fact that  $\ker(M(y)) = \text{Span}(y)$  as claimed.  $\square$

With this lemma established, it remains to show that  $e(f)$  defines a convex function. To see this, assume first that  $f \in \mathbb{R}_+^n$  is non-zero. We may then compute the gradient

$$\nabla e(f) = -(1 + \log\langle f, \mathbf{1} \rangle)\mathbf{1} + \sum_{i=1}^n \mathbf{e}_i \log\langle f, \mathbf{r}_i \rangle + \frac{f_i}{\langle f, \mathbf{r}_i \rangle} \mathbf{r}_i$$

as well as the Hessian

$$\text{Hess}_e(f) = -\frac{\mathbf{1}\mathbf{1}^T}{\langle f, \mathbf{1} \rangle} + \sum_{i=1}^n \frac{\mathbf{e}_i \mathbf{r}_i^T + \mathbf{r}_i \mathbf{e}_i^T}{\langle f, \mathbf{r}_i \rangle} - \mathbf{r}_i \frac{f_i}{\langle f, \mathbf{r}_i \rangle^2} \mathbf{r}_i^T$$

of the cluster energy. That  $\langle f, \mathbf{r}_i \rangle > 0$  for all  $1 \leq i \leq n$  follows from the assumption that  $P$  corresponds to a connected graph, so in particular all of the required derivatives exist. We may define a diagonal matrix  $\hat{F}$  with non-zero entries

$$\hat{F}_{ii} := \frac{1}{\langle f, \mathbf{r}_i \rangle} \quad \text{and} \quad \hat{F}^{-1} = \text{diag}(Qf)$$

that is well-defined, non-singular and positive-definite. We may then simplify the Hessian in matrix form as

$$\text{Hess}_e(f) = \hat{F}Q + Q^T \hat{F} - Q^T \hat{F} \text{diag}(f) \hat{F}Q - \frac{1}{\langle f, \mathbf{1} \rangle} \mathbf{1} \mathbf{1}^T,$$

with the aim in showing that  $\text{Hess}_e(f)$  has non-negative spectrum.

First, define  $z := Qf$  and  $y \in \text{int}(\mathbb{R}_+^n)$  via  $y_i = \sqrt{z_i}$  for each  $1 \leq i \leq n$ . Then take  $x \in \mathbb{R}^n$  arbitrary and write  $x = Q^{-1} \hat{F}^{-1} D_y^{-1} v$  to see

$$\langle x, \text{Hess}_e(f)x \rangle = \langle v, M(y)v \rangle$$

for  $M(y)$  and  $D_y^{-1}$  defined in lemma 11 above. Now write

$$x = \left( x - \frac{\langle x, \mathbf{1} \rangle}{\langle f, \mathbf{1} \rangle} f \right) + \frac{\langle x, \mathbf{1} \rangle}{\langle f, \mathbf{1} \rangle} f := x^0 + x^f,$$

and let  $v = v^0 + v^f$  denote the corresponding decomposition of  $v := D_y \hat{F} Q x$  after changing variables. As  $v^0 \perp y$  and  $v^f \in \text{Span}(y)$ , this yields

$$\begin{aligned} \langle x, \text{Hess}_e(f)x \rangle &= \langle x^0, \text{Hess}_e(f)x^0 \rangle \\ &= \langle v^0, M(y)v^0 \rangle = |v^0|_2^2 + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n P_{ij} \left( \frac{v_i^0 y_j}{y_i} - v_j^0 \right)^2 \end{aligned}$$

due to lemma 11. We may then use the fact that  $v_i^0 = y_i (\hat{F} Q x^0)_i$  to re-write the resulting expression

$$\langle x, \text{Hess}_e(f)x \rangle = \sum_{i=1}^n \frac{(Qx^0)_i^2}{\langle Qf, \mathbf{e}_i \rangle} + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n \left( \frac{(Qx^0)_i}{\langle Qf, \mathbf{e}_i \rangle} - \frac{(Qx^0)_j}{\langle Qf, \mathbf{e}_j \rangle} \right)^2 P_{ij} \langle Qf, \mathbf{e}_j \rangle \quad (46)$$

in the original variables.

That  $e(f)$  defines a convex function on  $\mathbb{R}_+^n$  now follows easily. Suppose first that  $f, g \in \mathbb{R}_+^n$  and that both  $f \neq \mathbf{0}$  and  $g \neq \mathbf{0}$  hold. Then for any  $t \in [0, 1]$  the linear interpolation

$$\ell_t := (1-t)f + tg \in \mathbb{R}_+^n$$

is non-zero as well, and so the cluster energy  $e(f)$  is twice differentiable along this line. For such  $f, g$  the identity

$$e(g) = e(f) + \langle \nabla e(f), g - f \rangle + \int_0^1 \langle g - f, \text{Hess}_e(f + t(g - f))(g - f) \rangle (1-t) dt \quad (47)$$

therefore holds. Now define

$$x_t := (g - f) - \frac{\langle g - f, \mathbf{1} \rangle}{\langle \ell_t, \mathbf{1} \rangle} \ell_t = \frac{1}{\langle \ell_t, \mathbf{1} \rangle} (\langle f, \mathbf{1} \rangle g - \langle g, \mathbf{1} \rangle f),$$

so that (46) yields  $\langle g - f, \text{Hess}_e(f + t(g - f))(g - f) \rangle =$

$$\sum_{i=1}^n \frac{(Qx_t)_i^2}{\langle Q\ell_t, \mathbf{e}_i \rangle} + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n \left( \frac{(Qx_t)_i}{\langle Q\ell_t, \mathbf{e}_i \rangle} - \frac{(Qx_t)_j}{\langle Q\ell_t, \mathbf{e}_j \rangle} \right)^2 P_{ij} \langle Q\ell_t, \mathbf{e}_j \rangle \quad (48)$$

for the inner product appearing in the integrand. Thus  $\langle g - f, \text{Hess}_e(f + t(g - f))(g - f) \rangle > 0$  unless  $x_t = 0$ , which occurs if and only if  $f$  and  $g$  are collinear. In particular, the strict inequality

$$e(g) > e(f) + \langle \nabla e(f), g - f \rangle \quad (49)$$

therefore holds for any pair  $f, g \in \mathbb{R}_+^n$  with  $f, g$  not collinear. This provides us with all of the ingredients necessary to prove



**Corollary 1** (Convexity and Strict Convexity) *Suppose  $f, g \in \mathbb{R}_+^n$  are non-zero and linearly independent. Define*

$$\ell_t := tg + (1-t)f \quad \text{and} \quad x_t := \frac{1}{\langle \ell_t, \mathbf{1} \rangle} (\langle f, \mathbf{1} \rangle g - \langle g, \mathbf{1} \rangle f).$$

*Then the equality*

$$\begin{aligned} e(g) = e(f) + \langle \nabla e(f), g - f \rangle + \sum_{i=1}^n \int_0^1 (1-t) \frac{(Qx_t)_i^2}{\langle Q\ell_t, \mathbf{e}_i \rangle} dt \\ + \frac{\alpha}{1-\alpha} \sum_{i=1}^n \sum_{j=1}^n \int_0^1 (1-t) \left( \frac{(Qx_t)_i}{\langle Q\ell_t, \mathbf{e}_i \rangle} - \frac{(Qx_t)_j}{\langle Q\ell_t, \mathbf{e}_j \rangle} \right)^2 P_{ij} \langle Q\ell_t, \mathbf{e}_j \rangle dt \end{aligned}$$

*holds, and in particular the strict inequality*

$$\theta e(g) + (1-\theta)e(f) > e(\theta f + (1-\theta)g) \quad (50)$$

*is valid for any  $0 < \theta < 1$  arbitrary. If  $f, g \in \mathbb{R}_+^n$  for  $f = \alpha g$  with  $\alpha \geq 0$  then*

$$\theta e(g) + (1-\theta)e(f) \geq e(\theta f + (1-\theta)g)$$

*for any  $0 \leq \theta \leq 1$ , and so  $e(f)$  defines a convex function on  $\mathbb{R}_+^n$ .*

**proof** A direct substitution of the equality (48) into (47) proves the first claim. To show (50), note that if  $0 < \theta < 1$  and  $f, g$  are not collinear then  $g$  and  $\theta g + (1-\theta)f$  are not collinear. Thus the strict inequality

$$e(g) > e(\theta g + (1-\theta)f) + (1-\theta)\langle \nabla e(\theta g + (1-\theta)f), g - f \rangle$$

holds by the first claim. Using  $f$  and  $\theta g + (1-\theta)f$  in the first claim also yields the symmetric inequality

$$e(f) > e(\theta g + (1-\theta)f) + \theta \langle \nabla e(\theta g + (1-\theta)f), f - g \rangle,$$

which yields (50) after adding  $\theta$  times the first inequality to  $(1-\theta)$  times the second inequality. Finally, suppose  $f, g \in \mathbb{R}_+^n$  and  $f = \alpha g$  for  $\alpha \geq 0$  a positive scalar. The one-homogeneity of  $e(f)$  then implies

$$\theta e(g) + (1-\theta)e(f) = \theta e(g) + (1-\theta)\alpha e(g) = e(\theta g + (1-\theta)\alpha g) = e(\theta g + (1-\theta)f),$$

which proves the final claim.  $\square$

## References

- [1] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/mccallum/bow>, 1996.
- [2] Zhirong Yang, Tele Hao, Onur Dikmen, Xi Chen, and Erkki Oja. Clustering by nonnegative matrix factorization using graph random walk. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1088–1096, 2012.
- [3] M. Stephane. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.