Online Supplementary Materials for "Positive Semidefinite Rank-based Correlation Matrix Estimation with Application to Semiparametric Graph Estimation"

Tuo Zhao[1] Kathryn Roeder[2] and Han Liu[3]

Department of Computer Science[1], Johns Hopkins University

Department of Statistics[2], Carnegie Mellon University

Department of Operations Research Research and Financial Engineering[3], Princeton University

S.1. Fast Projection Algorithm for elementwise ℓ_1-norm ball

By carefully examining (3.9), we find that it is a special case of the quadratic Knapsack problem (Brucker, 1984; Pardalos and Kovoor, 1990). We first define $\mathbf{A}' \in \mathbb{R}^{d\times d}$ with $\mathbf{A}'_{jk} = |\mathbf{A}|_{jk}/\mu$. Given the decreasing order statistics of all elements in $\mathbf{A}'$ as $\mathbf{A}'_{(1)}, \mathbf{A}'_{(2)}, ..., \mathbf{A}'_{(d^2)}$, (3.9) is equivalent to find u such that

$$\sum_{j=1}^{u-1}\left(\mathbf{A}'_{(j)} - \mathbf{A}'_{(u)}\right) < 1 \text{ and } \sum_{j=1}^{u+1}\left(\mathbf{A}'_{(j)} - \mathbf{A}'_{(u)}\right) \geq 1. \tag{S.1}$$

Then with u and $\mathbf{A}'_{(u)}$, we can calculate γ as follows,

$$\gamma = \frac{1}{u}\left(\sum_{j=1}^{u}\mathbf{A}'_{(j)} - 1\right). \tag{S.2}$$

Similar to the fast median algorithm (Corman et al., 2001), Algorithm 1 identifies u and the pivot value $\mathbf{A}'_{(u)}$ using a divide and conquer procedure (without sorting the data). In each iteration we either eliminate elements shown to be strictly smaller than $\mathbf{A}'_{(u)}$ or update the partial sum leading to (S.1). This algorithm has an average-case complexity of $O(d^2)$. Similar algorithms can be found in Liu and Ye (2009); Duchi et al. (2008) for the lasso problem.

Algorithm 1 The elementwise ℓ_1 norm projection algorithm.

Input: $\mathbf{A}' \in \mathbb{R}^{d\times d}$

Initialize: $\mathcal{S}_0 = \{(j,k) \mid j = 1,...,d \text{ and } k = 1,...,d\}$, $w = 0$,$u = 0$

repeat

1: randomly pick $(j',k') \in \mathcal{S}_0$

2: partition $\mathcal{S}_0$:

$$\mathcal{S}_1 = \{(j,k) \in \mathcal{S}_1 \mid \mathbf{A}'_{jk} \geq \mathbf{A}'_{j'k'}\}$$
$$\mathcal{S}_2 = \{(j,k) \in \mathcal{S}_2 \mid \mathbf{A}'_{jk} < \mathbf{A}'_{j'k'}\}$$

3: Calculate $\mathbf{\Delta}_w = |\mathcal{S}_1|$ and $\mathbf{\Delta}_u = \sum_{(j,k)\in\mathcal{S}_1} \mathbf{A}'_{jk}$

4: If $(u + \mathbf{\Delta}_u) - (w + \mathbf{\Delta}_w)\mathbf{A}'_{j'k'} \leq 1$

$u = u + \mathbf{\Delta}_u$; $w = w + \mathbf{\Delta}_w$; $\mathcal{S}_0 \leftarrow \mathcal{S}_2$

else

$\mathcal{S}_0 \leftarrow \mathcal{S}_1 \setminus \{(j',k')\}$

until $\mathcal{S}_0 = \emptyset$

Output: $\gamma = (w-1)/u$

S.1.1. Proof of Lemma 3.1

Proof. Equation (3.8) can be rewritten as

$$\vert\!\vert\!\vert \mathbf{A} \vert\!\vert\!\vert_\infty^\mu = \min_{\vert\!\vert\!\vert \mathbf{U} \vert\!\vert\!\vert_1 \leq 1} \frac{\mu}{2} ||\mathbf{U} - \mathbf{A}/\mu||_{\mathrm{F}}^2. \tag{S.3}$$

By the Lagrangian duality, we know that there exists some constant $\gamma > 0$ such that

$$\vert\!\vert\!\vert \mathbf{A} \vert\!\vert\!\vert_\infty^\mu = \min_{\mathbf{U}} ||\mathbf{U} - \mathbf{A}/\mu||_{\mathrm{F}}^2 + \gamma \vert\!\vert\!\vert \mathbf{U} \vert\!\vert\!\vert_1 \tag{S.4}$$

holds as a Lagrangian form equivalent to (S.3). (S.4) results in the soft thresholding operation as follow

$$\widetilde{\mathbf{U}}_{jk} = \mathrm{sign}\left(\mathbf{A}_{jk}\right) \cdot \max\left\{\left|\frac{\mathbf{A}_{jk}}{\mu}\right| - \gamma, 0\right\}, \tag{S.5}$$

which completes the proof. $\square$

S.2. Accelerated Proximal Gradient Algorithm

We summarize the accelerated proximal gradient algorithm in Algorithm 2.

S.3. Proof of Lemma 3.2

Proof. The eigenvalue decomposition of $\mathbf{A}$ can be rewritten as $\mathbf{A} = \mathbf{V}\mathbf{Z}\mathbf{V}^T$ with

$$\mathbf{Z} = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_d) \text{ and } \mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_d). \tag{S.6}$$

Algorithm 2 The Accelerated Proximal Gradient Algorithm.

Input: $\widehat{\mathbf{S}}$, $\mathbf{S}^{(0)} = \mathbf{M}^{(0)} = \mathbf{W}^{(0)}$, μ, $\theta_t = 2/(1+t)$, ε
Output: $\widetilde{\mathbf{S}} = \mathbf{S}^{(t)}$
Initialize: $t = 1$
repeat
1: Compute the auxiliary variables $\mathbf{M}^{(t)}$ using (3.13)
2: Compute the gradient of (3.10) at $\mathbf{M}^{(t)}$ using (3.14)
3: (Optional) Compute η_t by the backtracking line search using (3.18)
4: Compute the auxiliary variables $\mathbf{W}^{(t)}$ using (3.16)
5: Compute the solution $\mathbf{S}^{(t)}$ using (3.17)
6: $t = t + 1$
until $\left| \|\|\widehat{\mathbf{S}} - \mathbf{S}^{(t)}\|\|_{\infty}^{\mu} - \|\|\widehat{\mathbf{S}} - \mathbf{S}^{(t-1)}\|\|_{\infty}^{\mu} \right| \le \varepsilon\mu.$

Note is that $\mathbf{V}$ is a unitary matrix. Since the Frobenius norm is invariant to $\mathbf{V}$, we have

$$\min_{\mathbf{B}\succeq 0} ||\mathbf{B} - \mathbf{A}||_{\mathrm{F}}^2 = \min_{\mathbf{B}\succeq 0} ||\mathbf{V}^T(\mathbf{B} - \mathbf{A})\mathbf{V}||_{\mathrm{F}}^2 = \min_{\mathbf{B}\succeq 0} ||\mathbf{V}^T\mathbf{B}\mathbf{V}^T - \mathbf{Z}||_{\mathrm{F}}^2. \tag{S.7}$$

Then it is easy to verify that (S.7) is minimized when

$$\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{R} = \mathrm{diag}(\widetilde{\sigma}_1, \widetilde{\sigma}_2, ..., \widetilde{\sigma}_d), \tag{S.8}$$

where $\widetilde{\sigma}_j = \max\{\sigma_j, 0\}$. Therefore we have

$$\mathbf{B} = \mathbf{V}\mathbf{R}\mathbf{V}^T, \tag{S.9}$$

which completes the proof. □

S.4. Proof of Lemmas related to Theorem 4.2

S.4.1. Proof of Lemma C.1

Proof.

$$\begin{aligned}
\|\mathbf{AB}\|_\infty &= \max_i \sum_j \sum_k |\mathbf{A}_{ik}\mathbf{B}_{kj}| = \max_i \sum_k |\mathbf{A}_{ik}| \sum_j |\mathbf{B}_{kj}| \\
&\le \left(\max_i \sum_k |\mathbf{A}_{ik}|\right)\left(\max_\ell \sum_j |\mathbf{B}_{\ell j}|\right) = \|\mathbf{A}\|_\infty \|\mathbf{B}\|_\infty.
\end{aligned} \tag{S.10}$$

□

S.4.2. Proof of Lemma C.2

Proof.

$$\begin{aligned}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty &= \|\widehat{\mathbf{B}}^{-1}(\mathbf{B}-\widehat{\mathbf{B}})\mathbf{B}^{-1})\|_\infty \le \|\widehat{\mathbf{B}}^{-1}\|_\infty\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty \\ &\le \left(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\|\mathbf{B}^{-1}\|_\infty\right)\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty. \end{aligned} \tag{S.11}$$

Therefore we have

$$\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty \le \frac{\|\mathbf{B}^{-1}\|_\infty^2}{1-\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty}\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty, \tag{S.12}$$

which completes the proof. □

S.4.3. Proof of Lemma C.3

Proof. We have the following decomposition,

$$\left\{(\widehat{\mathbf{A}}-\mathbf{A})(\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1})\right\} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}^{-1}-\mathbf{A}\mathbf{B}^{-1}+\left\{(\widehat{\mathbf{A}}-\mathbf{A})\mathbf{B}^{-1}\right\}+\left\{\mathbf{A}(\mathbf{B}^{-1}-\widehat{\mathbf{B}}^{-1})\right\}. \tag{S.13}$$

Since

$$\begin{aligned}\left\{\mathbf{A}(\mathbf{B}^{-1}-\widehat{\mathbf{B}}^{-1})\right\} &= \mathbf{A}\mathbf{B}^{-1}(\mathbf{I}-\mathbf{B}\widehat{\mathbf{B}}^{-1}) = \mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})\widehat{\mathbf{B}}^{-1} \\ &= \mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})(\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1})+\mathbf{A}\mathbf{B}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{B}^{-1}, \end{aligned} \tag{S.14}$$

then we have

$$\begin{aligned}\|\widehat{\mathbf{A}}\widehat{\mathbf{B}}^{-1}-\mathbf{A}\mathbf{B}^{-1}\|_\infty &\le \left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty\right\}+\left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\psi\right\} \\ &\quad +\left\{\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty\right\}+\left\{\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\psi\right\} \\ &= \left\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\psi)+(\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\psi)\right. \\ &= (\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty+\alpha\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty)(\psi+\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty). \end{aligned} \tag{S.15}$$

By (S.11) in Lemma C.2

$$\begin{aligned}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty &\le \left(\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\|\mathbf{B}^{-1}\|_\infty\right)\|\mathbf{B}-\widehat{\mathbf{B}}\|_\infty\|\mathbf{B}^{-1}\|_\infty \\ &= \frac{1}{2}\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty+\frac{1}{2}\psi, \end{aligned} \tag{S.16}$$

we have

$$\|\widehat{\mathbf{B}}^{-1}-\mathbf{B}^{-1}\|_\infty \le \psi. \tag{S.17}$$

By combing (S.15) and (S.17), we complete the proof. □

S.4.4. Proof of Lemma C.4

Proof.

$$\begin{aligned}\|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}} - \mathbf{A}\boldsymbol{v}\|_\infty &= \|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}} - \mathbf{A}\widehat{\boldsymbol{v}} + \mathbf{A}\widehat{\boldsymbol{v}} - \mathbf{A}\boldsymbol{v}\|_\infty \\ &\leq \|\widehat{\mathbf{A}}\widehat{\boldsymbol{v}} - \mathbf{A}\widehat{\boldsymbol{v}}\|_\infty + \|\mathbf{A}\widehat{\boldsymbol{v}} - \mathbf{A}\boldsymbol{v}\|_\infty \\ &\leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty(\|\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_\infty + \|\boldsymbol{v}\|_\infty) + \|\mathbf{A}\|_\infty\|\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_\infty, \end{aligned} \tag{S.18}$$

which completes the proof. □

S.4.5. Proof of Lemma C.5

Proof. We have

$$\begin{aligned} ||\widetilde{\mathbf{S}}_{I_jI_j} - \boldsymbol{\Sigma}^*_{I_jI_j}||_\infty &= \max_{k\in I_j}\sum_{\ell\in I_j}|\widetilde{\mathbf{S}}_{k\ell} - \boldsymbol{\Sigma}^*_{k\ell}| \leq \kappa_3 s\sqrt{\frac{\log d}{n}}, \\ ||\widetilde{\mathbf{S}}_{J_jI_j} - \boldsymbol{\Sigma}^*_{J_jI_j}||_\infty &= \max_{k\in J_j}\sum_{\ell\in I_j}|\widetilde{\mathbf{S}}_{k\ell} - \boldsymbol{\Sigma}^*_{k\ell}| \leq \kappa_3 s\sqrt{\frac{\log d}{n}}, \\ ||\widetilde{\mathbf{S}}_{\setminus j,j} - \boldsymbol{\Sigma}^*_{\setminus j,j}||_\infty &= \max_{k\neq j}|\widetilde{\mathbf{S}}_{kj} - \boldsymbol{\Sigma}^*_{kj}| \leq \kappa_3\sqrt{\frac{\log d}{n}}. \end{aligned}$$

Let $s_j = |I_j| \leq s$. For arbitrary $\boldsymbol{v} \in \mathbb{R}^{s_j}$, we have

$$\begin{aligned} \boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} &= \boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j} - \boldsymbol{\Sigma}^*_{I_jI_j} + \boldsymbol{\Sigma}^*_{I_jI_j})\boldsymbol{v} \\ &= \boldsymbol{v}^T\boldsymbol{\Sigma}^*_{I_jI_j}\boldsymbol{v} - \boldsymbol{v}^T(\boldsymbol{\Sigma}^*_{I_jI_j} - \widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} \\ &\geq \Lambda_{\min}(\boldsymbol{\Sigma}^*_{I_jI_j})||\boldsymbol{v}||_2^2 - ||\boldsymbol{v}||_1^2 \big\|\big\|\boldsymbol{\Sigma}^*_{I_jI_j} - \widetilde{\mathbf{S}}_{I_jI_j}\big\|\big\|_\infty \\ &\geq \delta||\boldsymbol{v}^T||_2^2 - s_j||\boldsymbol{v}||_2^2 \cdot \kappa_3\sqrt{\frac{\log d}{n}}, \end{aligned}$$

where the last inequality comes from the fact $\boldsymbol{v} \in \mathbb{R}^{s_j}$. Thus for large enough n such that

$$\sqrt{\frac{\log d}{n}} \leq \frac{\delta}{2s\kappa_3}, \tag{S.19}$$

we have

$$\boldsymbol{v}^T(\widetilde{\mathbf{S}}_{I_jI_j})\boldsymbol{v} \geq \frac{\delta}{2}||\boldsymbol{v}^T||_2^2. \tag{S.20}$$

Since $\boldsymbol{v}$ and j are arbitrary, we further have

$$\Lambda_{\min}(\widetilde{\mathbf{S}}_{I_jI_j}) \geq \frac{\delta}{2} \text{ for all } j = 1, ..., d. \tag{S.21}$$

□

S.5. Graph and Covariance Matrix Generation

• **Neighborhood**. For each node, we independently sample a random vector from a uniform distribution over $[0,1]^2 \subset \mathbb{R}^2$. Let $\boldsymbol{V}_i \in \mathbb{R}^2$ denote the random vector for node i, then we set an edge between node i and node j with probability $(2\pi)^{-1/2}\exp(-||\boldsymbol{V}_i - \boldsymbol{V}_j||_2^2/\phi)$. We set $\phi = 100$ for the simulations in §5.1, §5.2, and §5.3.

• **Clique**. The nodes are evenly partitioned into g disjoint groups and each group contains d/g nodes. The subgraph of each group is fully connected graph. We set $g = 20$ for the simulations in §5.1, §5.2, and §5.3.

• **Band**. Each node is assigned a coordinate j with $j = 1, ..., d$. Two nodes are connected by an edge whenever the corresponding points are at distance no more than g. We set $g = 2$ for the simulations in §5.1, §5.2, and §5.3.

• **Lattice**. Each node is assigned a two dimensional coordinate (j,k) with $j = 1, ..., g$ and $k = 1, ..., d/g$. Two nodes are connected by an edge whenever the corresponding points are at distance 1. We set $g = 10$ for the simulations in §5.1, §5.2, and §5.3.

• **Mixed Scale-free**. The nodes are evenly partitioned into 4 groups. The nodes from different groups are disconnected. The subgraph of each group of nodes is a scale free graph. The degree distribution of the scale-free graph follows a power law. The graph is generated by the preferential attachment mechanism. The graph begins with an initial band graph of 10 nodes with $g = 1$. New nodes are added to the graph one at a time. Each new node is connected to existing node with a probability that is proportional to the number of degrees that the existing nodes already have. Formally, the probability p_i that the new node is connected to node i is, $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is the degree of node i.

• **Hybrid**. The nodes are evenly partitioned in to 5 groups, named S_1-S_5. The subgraph of S_1 is a neighborhood graph with $\phi = 25$; The subgraph of S_2 is a clique graph with $g = 4$; The subgraph of S_3 is a band graph with $g = 2$; The subgraph of S_4 is a lattice graph with $g = 10$; The subgraph of S_5 is a scale-free graph. In addition, we set an edge between a node in S_k and a node in S_{k+1} with probability 0.01, independently of the other edges for $k = 1, ..., 4$.

Recall that $\mathbf{E}^*$ denotes the binary adjacency matrix, we calculate

$$\boldsymbol{\Sigma}^* = \mathcal{C}_2\{(\mathbf{E}^* + (0.5 - \Lambda_{\min}(\mathbf{E}^*)) \cdot \mathbf{I}_{200})^{-1}\}, \tag{S.22}$$

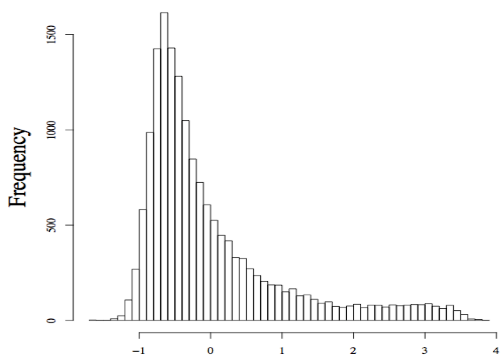
where $\mathcal{C}_2$ is the rescaling operator that converts a covariance matrix to the corresponding correlation matrix.

S.6. Supporting Figures for Data Analysis

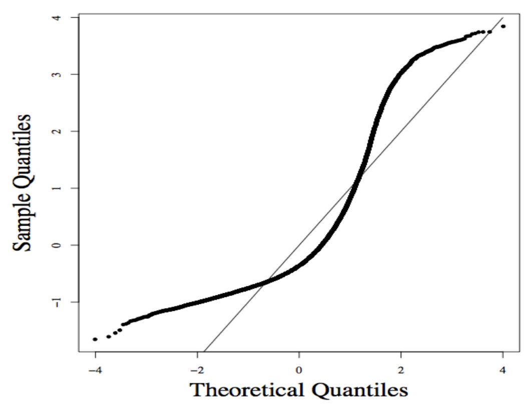
Due to the space limit of the manuscript, we present some supporting figures here.

S.6.1. Topic Modeling Dataset

Figure S.1 shows the histogram and normal qq plot of Topic 4. We see that Topic 4 significantly violates of the normality assumption.



(a) Histogram



(b) Normal qq plot

FIG S.1. *Both the histogram and normal qq plot show significant violation of the normality.*

The modules of the nonparanormal graph are magnified and presented in Figures S.2-S.5. The modules of the Gaussian graph are magnified and presented in Figures S.6-S.8

S.6.2. Stock Market Dataset

We plot the data points for the first 100 stocks in Figure S.9. We highlight a data point in red if its absolute value is greater than 3. We can see that a large number of potential outliers exist. They may affect the quality of the estimated graph.

S.6.3. Gene Network Dataset

Figure S.10 shows the histogram and normal qq plot of "MECPS". We see that its distribution is very likely non-Guassian.

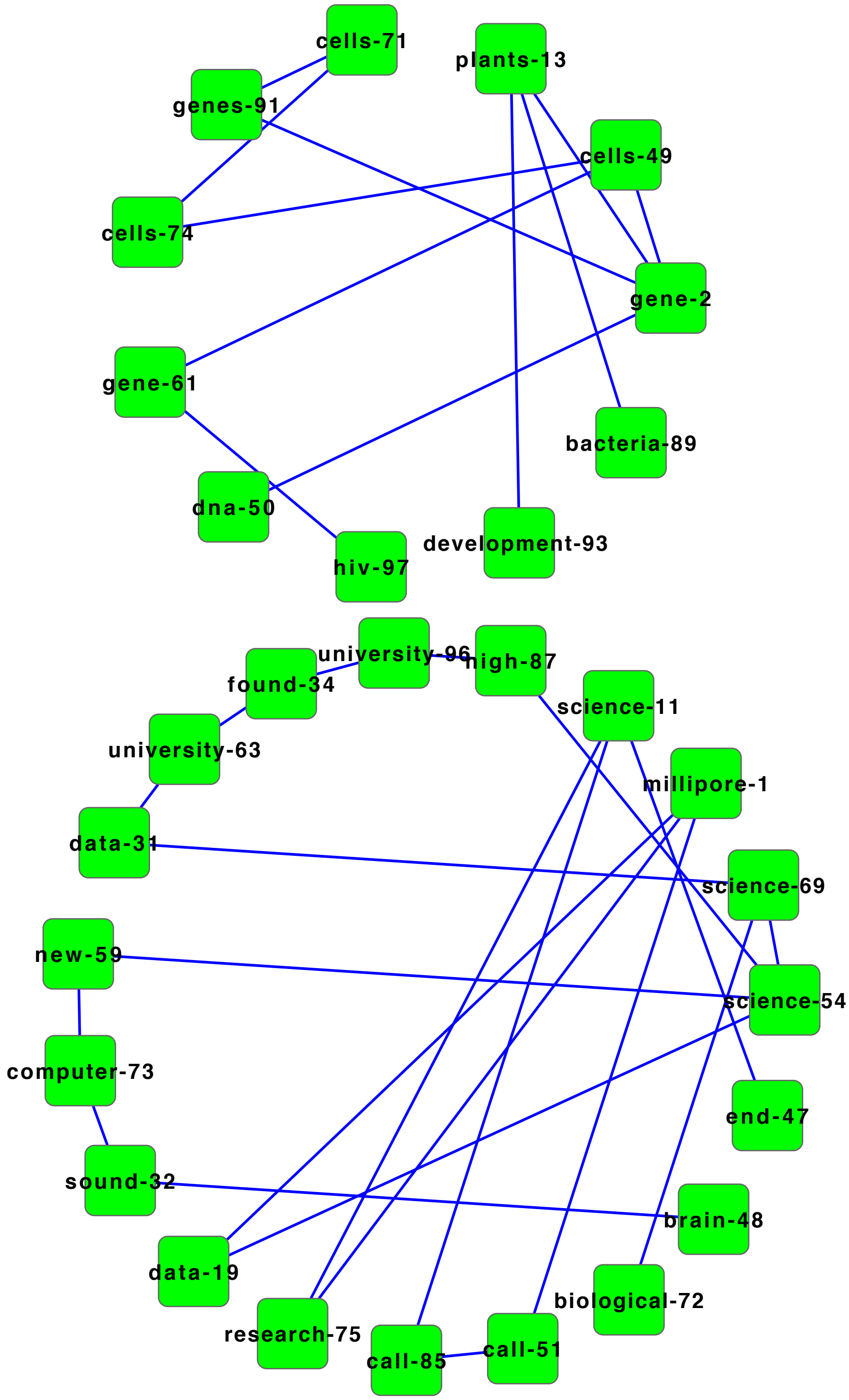


FIG S.2. *The nonparanormal topic graph. Part 1.*

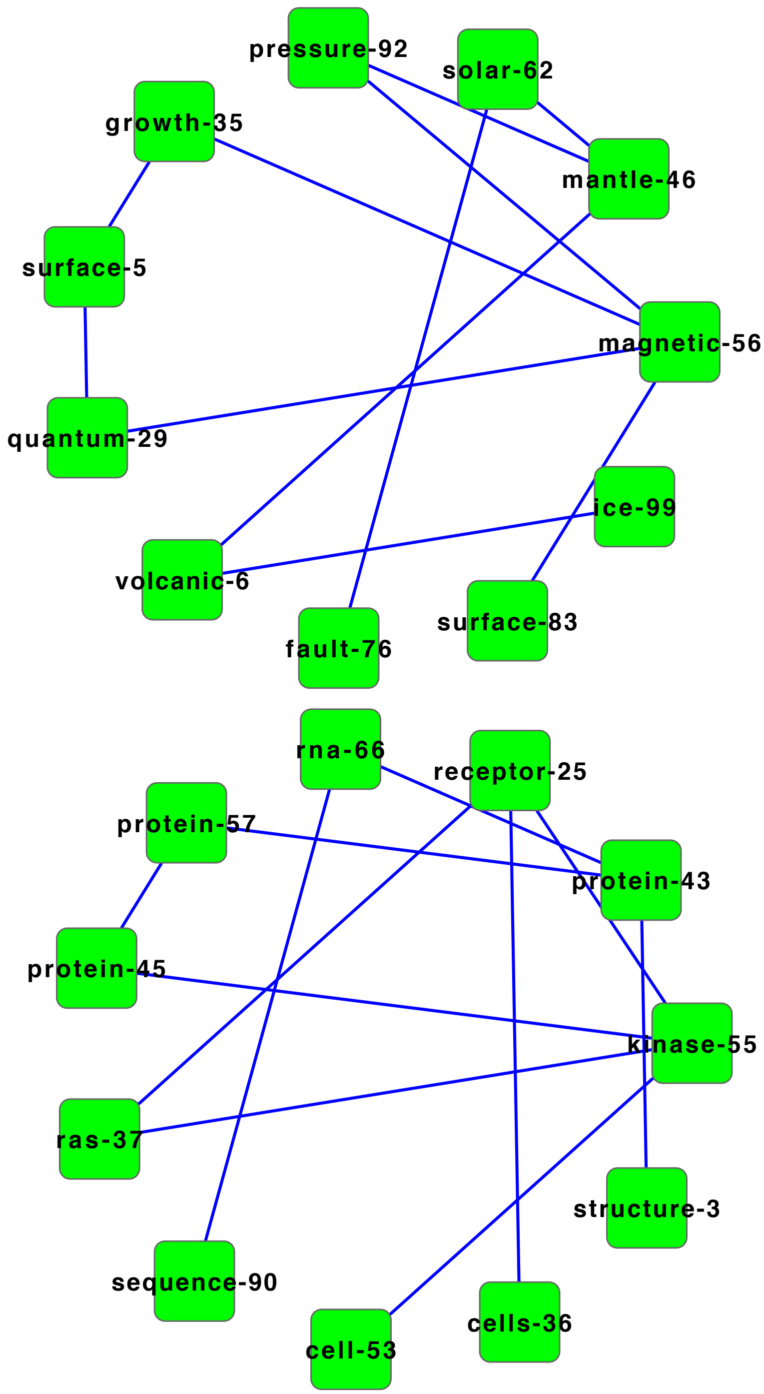


Fig S.3. *The nonparanormal topic graph. Part 2.*

References

Brucker, P. (1984). An O(n) Algorithm for Quadratic Knapsack Problems. *Operations Research*

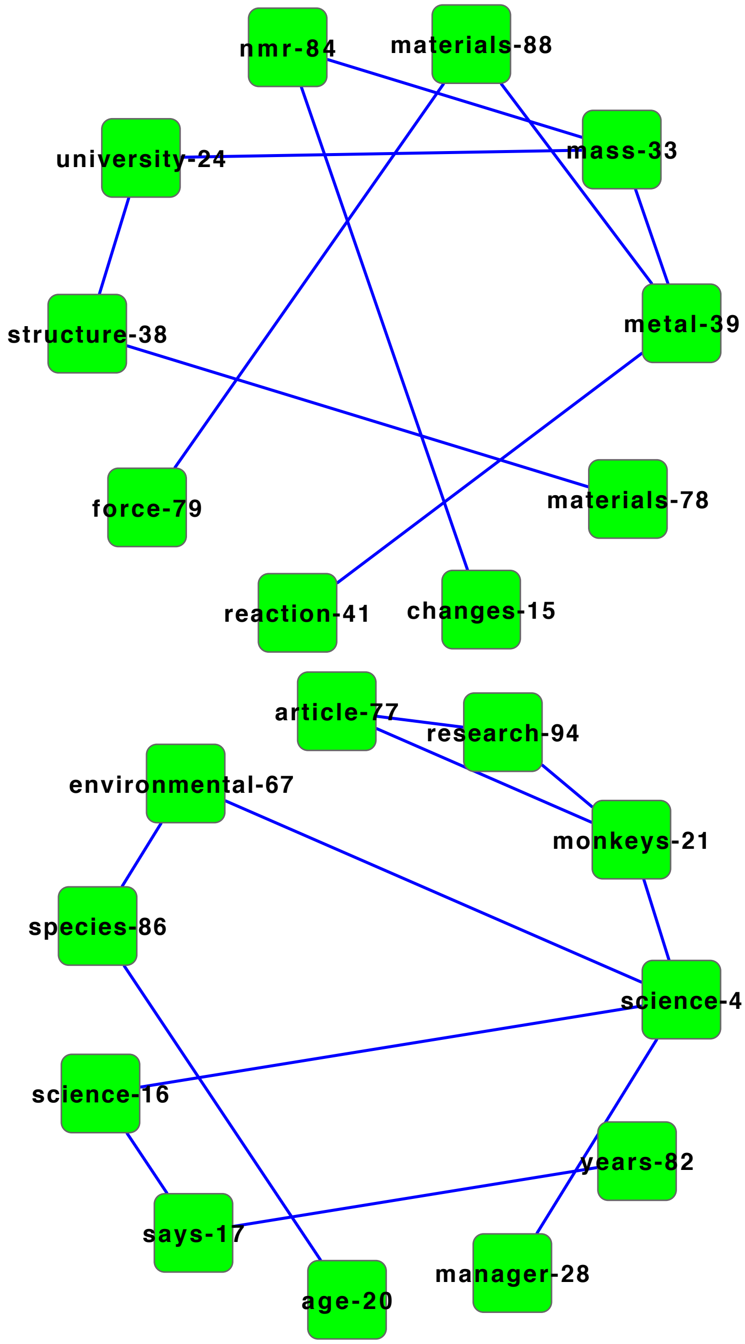


FIG S.4. *The nonparanormal topic graph. Part 3.*

Letters **3** 163-166.

CORMAN, T., LEISERSON, C., RIVEST, R. and STEIN, C. (2001). *Introduction to algorithms*. MIT

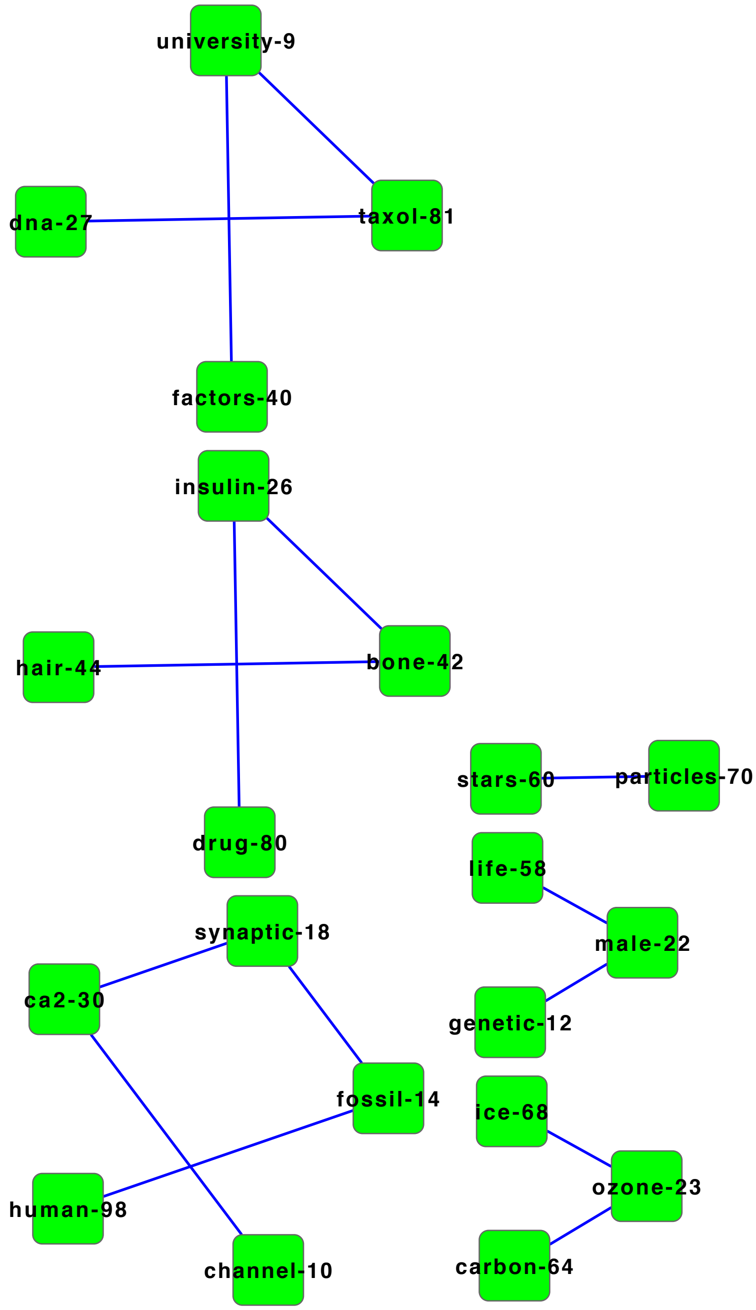


FIG S.5. *The nonparanormal topic graph. Part 4.*

Press.

DUCHI, J., SHWARTZ, S., SINGER, Y. and CHANDRA, T. (2008). Efficient projections onto the L1-ball for learning in high dimensions. *International Conference on Machine Learning* 272-279.

LIU, J. and YE, J. (2009). Efficient Euclidean Projections in Linear Time. *International Conference on Machine Learning*.

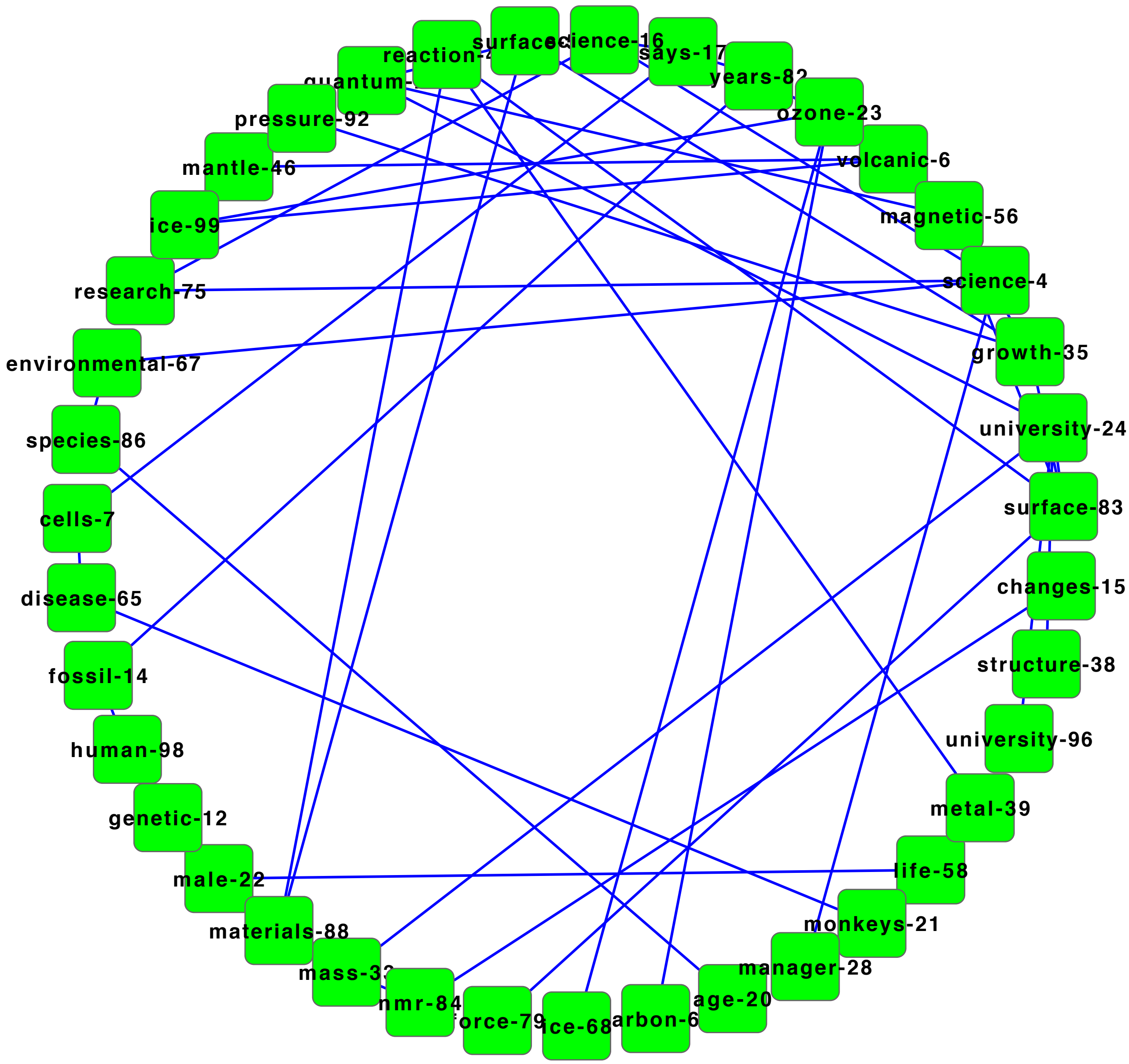


Fig S.6. *The Gaussian topic graph. Part 1.*

Pardalos, P. M. and Kovoor, N. (1990). An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. *Mathematical Programming* **46** 312-328.

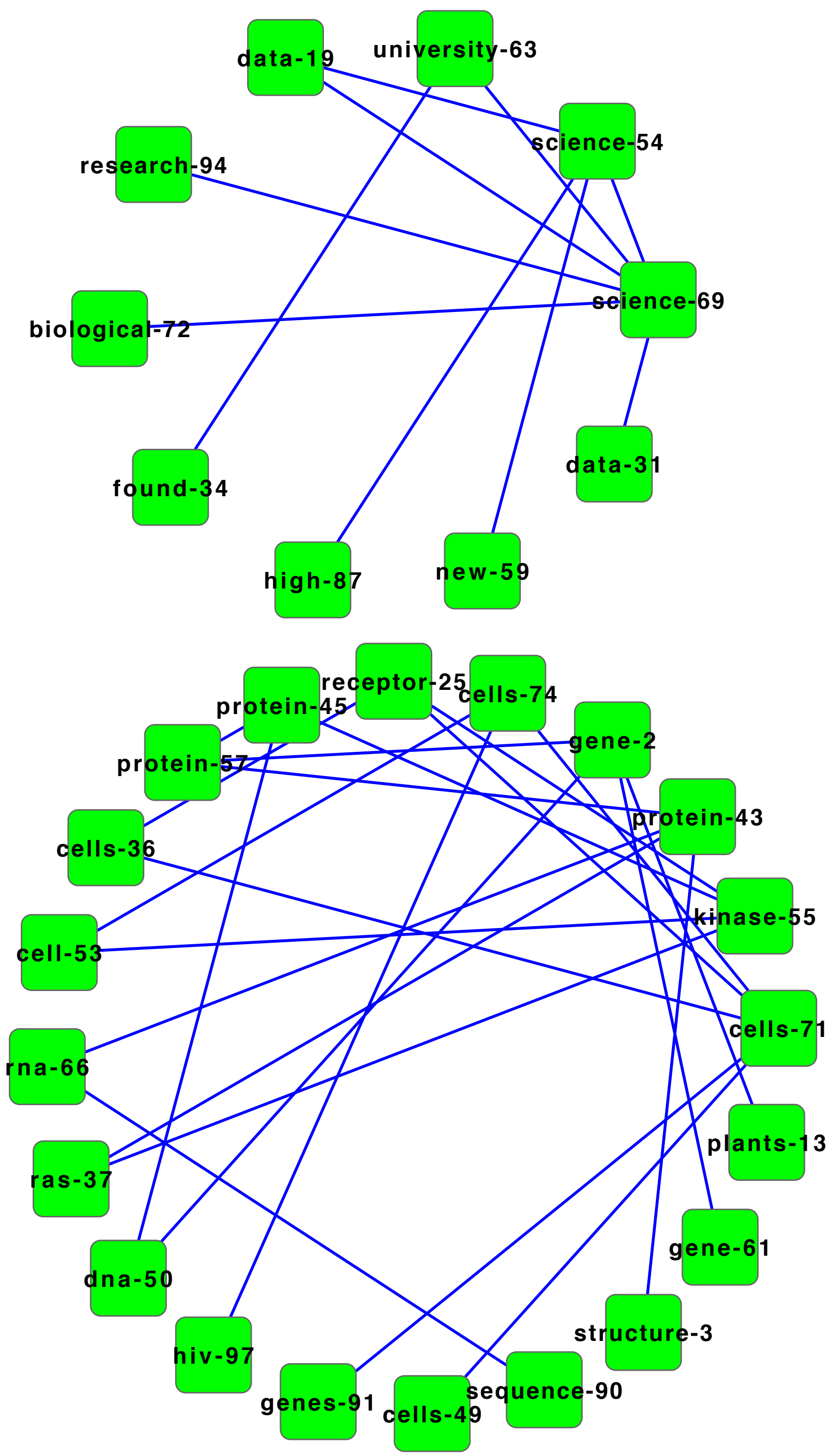


FIG S.7. *The Gaussian topic graph. Part 2.*

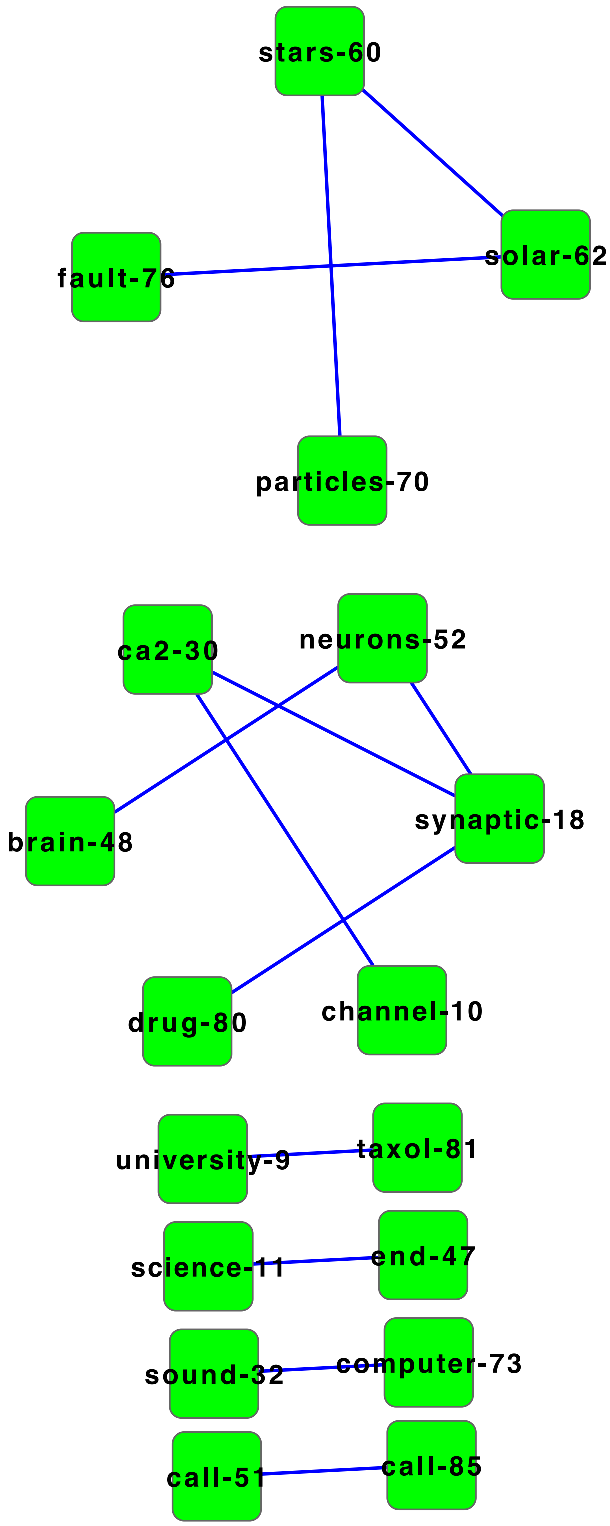


FIG S.8. *The Gaussian topic graph. Part 3.*

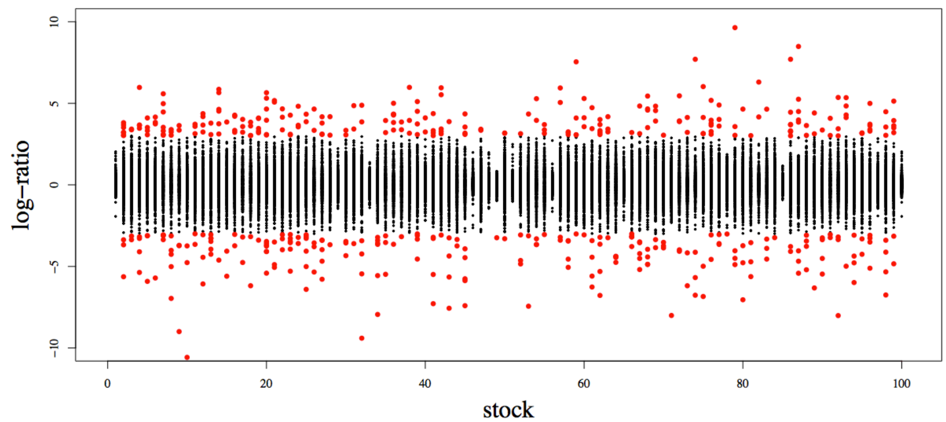


Fig S.9. *Stock Market Dataset. We can see a large amounted of the outliers (Red dots). Their existence may affect the quality of the estimated graph.*

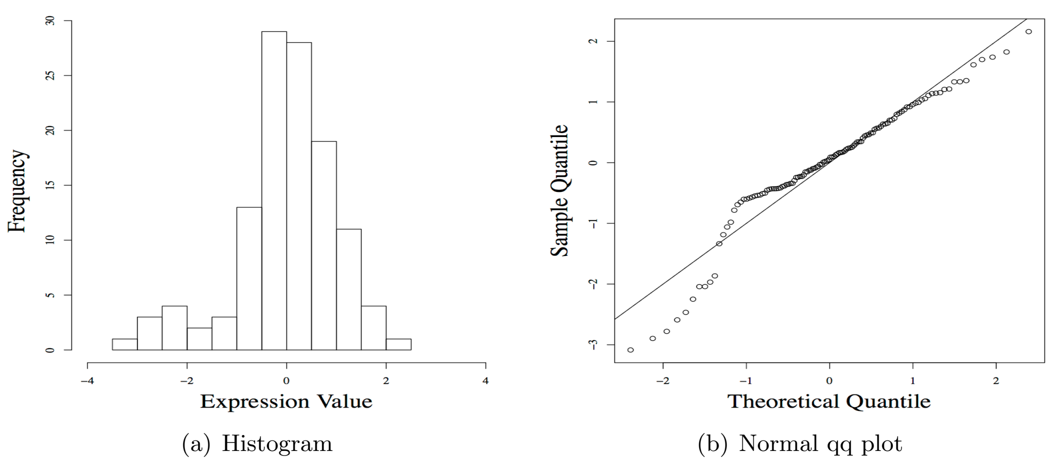


Fig S.10. *Both the histogram and normal qq plot show violation of the normality.*