Optimization Equivalence of Divergences Improves Neighbor Embedding (Supplemental Document)

Zhirong Yang[2] ZHIRONG.YANG@AALTO.FI
Jaakko Peltonen[1,4] JAAKKO.PELTONEN@AALTO.FI
Samuel Kaski[1,3] SAMUEL.KASKI@AALTO.FI

[1]Helsinki Institute for Information Technology HIIT, [2]Department of Information and Computer Science, Aalto University, Finland, [3]Department of Computer Science, University of Helsinki, and [4]University of Tampere

This supplemental document is organized as follows. Section 1 gives the source and brief descriptions of the datasets used in the paper. Section 2 gives a detailed information retrieval interpretation of ws-SNE. Section 3 provides additional visualizations that do not fit within the page limit of the main paper.

1. Datasets

We present results of six datasets in the paper.

- `shuttle`: the data is from the UCI Machine Learning Repository, Statlog (Shuttle) Data Set[1]. It contains 58000 samples from 7 classes, where each sample has 9 numerical shuttle attributes.
- `MNIST`: the data is from the MNIST database[2]. It contains 70000 handwritten digit images from 10 classes. We preprocessed the images by the scattering operator (Mallat, 2012) and PCA, which yields 256 features for each sample.
- `worldtrade`: the data is from Pajek datasets[3]. It is a weighted graph whose edges are trading amounts between 80 countries in the world.
- `usair97`: the data is from the LinLog layout package[4]. It is a binary graph where the nodes are 332 airports in the United States and the edges indicate whether there is direct flight between the airports.
- `mirex07`: the data is from the the Third Music Information Retrieval Evaluation eXchange (MIREX 2007)[5]. We used the version from the collection by Chen et al. (2009). It is a similarity graph of 3090 songs. The songs are evenly divided among 10 classes that roughly correspond to different music genres. The weighted edges are human judgment on how similar two songs are.
- `luxembourg`: the data is from The University of Florida Sparse Matrix Collection[6]. It contains 114599 nodes and 239332 edges, where each edge is a street in Luxembourg.

To maintain the space limit in the paper, we present results of two other datasets in this supplemental document (see Section 3):

- `jazz`: the data is from Arenas's collection[7]. It is a social network of 198 musicians. The musicians are mainly in New York and Chicago, with a few in both places or in other places.
- `ca-GrQc`: the data is from Stanford Network Analysis Project[8]. It is a coauthor network among 5242 researchers working on general relativity. We extracted the largest connected component with 4158 authors.

2. Retrieval interpretation of ws-SNE

Consider the objective of ws-SNE, $\mathcal{J}_{\text{ws-SNE}}(Y) = D_{\text{KL}}(p||M \circ q)$, where $M_{ij} = d_i d_j$ and d_i represents an importance of node i which is known or computable from the data, such as number of neighbors of the node (below this is also denoted as 'high degree'). We will show $\mathcal{J}_{\text{ws-SNE}}(Y)$ has an information retrieval interpretation as maximizing recall of retrieved neighbors, weighted by severity of the left-out misses.

[1] http://archive.ics.uci.edu/ml/
[2] http://yann.lecun.com/exdb/mnist/
[3] http://vlado.fmf.uni-lj.si/pub/networks/data/
[4] http://code.google.com/p/linloglayout/
[5] http://www.music-ir.org/mirex/wiki/2007
[6] http://www.cise.ufl.edu/research/sparse/matrices/
[7] http://deim.urv.cat/~aarenas/data/welcome.htm
[8] http://snap.stanford.edu/index.html

Notation. As in the main paper, let $\tilde{p}_{ij} = \frac{p_{ij}}{\sum_{kl} p_{kl}}$ be a symmetric input distribution jointly over nodes and their neighbors in the original space. Here p_{ij} are symmetric but not necessarily normalized, and the normalization in $\tilde{p}_{ij}$ ensures $\sum_{ij} \tilde{p}_{ij} = 1$. Let $\tilde{q}_{ij} = \frac{M_{ij} q_{ij}}{\sum_{kl} M_{kl} q_{kl}} = \frac{d_i d_j q_{ij}}{\sum_{kl} d_k d_l q_{kl}}$ be a distribution over nodes and their neighbors on the display such that q_{ij} is Gaussian $q_{ij} = \exp(-||\mathbf{y}_i - \mathbf{y}_j||^2)$ or Cauchy $q_{ij} = (1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}$. Note that again the q_{ij} are not normalized, and the normalization in $\tilde{q}_{ij}$ ensures $\sum_{ij} \tilde{q}_{ij} = 1$.

The joint distribution $\tilde{q}_{ij}$ depends on locations of nodes on the display and on their importances; intuitively, this distribution represents retrieval behavior of an analyst who looks at the display in a quick manner, noticing high-importance nodes and close-by other nodes, and the task of the visualization is to ensure that such retrieval of nodes and neighbors corresponds to the true neighbors represented in $\tilde{p}_{ij}$.

Each joint distribution over nodes can be written as a product of a marginal and a conditional distribution, so that $\tilde{p}_{ij} = \tilde{p}_i \tilde{p}_{j|i}$ where $\tilde{p}_i = \sum_k \tilde{p}_{ik} = \sum_k p_{ik} / \sum_{lm} p_{lm}$ and $\tilde{p}_{j|i} = \tilde{p}_{ij} / \tilde{p}_i = p_{ij} / \sum_{ik} p_{ik}$, and similarly $\tilde{q}_{ij} = \tilde{q}_i \tilde{q}_{j|i}$ where

$$\tilde{q}_i = \sum_k \tilde{q}_{ik} = \frac{d_i \sum_k d_k q_{ik}}{\sum_{kl} d_k d_l q_{kl}} \quad \text{and} \quad \tilde{q}_{j|i} = \frac{\tilde{q}_{ij}}{\tilde{q}_i} = \frac{d_j q_{ij}}{\sum_k d_k q_{ik}}\ . \tag{1}$$

Note that $\sum_i \tilde{p}_i = 1$ and $\sum_j \tilde{p}_{j|i} = 1$ for all i, and similarly $\sum_i \tilde{q}_i = 1$ and $\sum_j \tilde{q}_{j|i} = 1$ for all i.

Inserting the above into the cost function of ws-SNE, the Kullback-Leibler divergence becomes a sum of two terms, a divergence between the marginals and a weighted mean of divergences between conditionals:

$$\begin{aligned}
&\mathcal{J}_{\text{ws-SNE}}(Y) && (2)\\
=&D_{\text{KL}}(p||M \circ q) && (3)\\
=&\sum_{ij} \Big(\frac{p_{ij}}{\sum_{kl} p_{kl}}\Big) \log \frac{\left(\frac{p_{ij}}{\sum_{kl} p_{kl}}\right)}{\left(\frac{d_i d_j q_{ij}}{\sum_{kl} d_k d_l q_{kl}}\right)} && (4)\\
=&\sum_{ij} \tilde{p}_{ij} \log \frac{\tilde{p}_{ij}}{\tilde{q}_{ij}} && (5)\\
=&\sum_{ij} \tilde{p}_i \tilde{p}_{j|i} \log \frac{\tilde{p}_i \tilde{p}_{j|i}}{\tilde{q}_i \tilde{q}_{j|i}} && (6)\\
=&\sum_{ij} \tilde{p}_i \tilde{p}_{j|i} \Big(\log \frac{\tilde{p}_i}{\tilde{q}_i} + \log \frac{\tilde{p}_{j|i}}{\tilde{q}_{j|i}} \Big) && (7)\\
=&\sum_i \tilde{p}_i \Big(\sum_j \tilde{p}_{j|i} \Big) \log \frac{\tilde{p}_i}{\tilde{q}_i} + \sum_{ij} \tilde{p}_i \tilde{p}_{j|i} \log \frac{\tilde{p}_{j|i}}{\tilde{q}_{j|i}} && (8)\\
=&\sum_i \tilde{p}_i \log \frac{\tilde{p}_i}{\tilde{q}_i} + \sum_i \tilde{p}_i \sum_j \tilde{p}_{j|i} \log \frac{\tilde{p}_{j|i}}{\tilde{q}_{j|i}} && (9)\\
=&D_{\text{KL}}(\{\tilde{p}_i\}||\{\tilde{q}_i\}) + \sum_i \tilde{p}_i D_{\text{KL}}(\{\tilde{p}_{j|i}\}||\{\tilde{q}_{j|i}\}) && (10)
\end{aligned}$$

where $\{\tilde{p}_i\}$ is the distribution formed by all values $\tilde{p}_i$, and similarly for $\{\tilde{q}_i\}$; and $\{\tilde{p}_{j|i}\}$ is the distribution formed by all values $\tilde{p}_{j|i}$ for a fixed i, and similarly for $\{\tilde{q}_{j|i}\}$.

We next analyze the conditional divergences and provide an information retrieval interpretation for them, and then analyze the marginal divergence and provide an information retrieval interpretation for it.

2.1. Analysis of the conditional divergences

We first analyze the second term which is a weighted average of Kullback-Leibler divergences between conditional distributions of neighbors. Each Kullback-Leibler divergence in the second term compares the true neighborhood of node i to an on-screen neighborhood where users are likely to retrieve close-by nodes, or nodes with high degree, and counts the cost of misses in such retrieval. We now show this has an information retrieval interpretation.

First, note that the weighted conditional output probabilities $\tilde{q}_{j|i}$ in (1) can be written as

$$\tilde{q}_{j|i} = \frac{d_j q_{ij}}{\sum_k d_k q_{ik}} = \frac{d_j \frac{q_{ij}}{\sum_l q_{il}}}{\sum_k d_k \frac{q_{ik}}{\sum_l q_{il}}} = \frac{d_j \frac{q_{ij}}{\sum_l q_{il}}}{G_i}\ . \tag{11}$$

where the $q_{ij} / \sum_l q_{il}$ are unweighted conditional output probabilities, which are based only on how close points are to points i on the display but not on the importances of the points. Here we denoted the denominator by $G_i = \sum_k d_k \frac{q_{ik}}{\sum_l q_{il}}$ which is the weighted average importance of nodes near i on the display, where the importance of each node k is weighted by $\frac{q_{ik}}{\sum_l q_{il}}$.

Analysis in a simplified situation. Consider a simplified situation where the input probability $\tilde{p}_{j|i}$ takes a high value $A_i = \frac{1-\delta}{R_i}$ for R_i points and a low value $B_i = \frac{\delta}{N-R_i-1}$ for other points, where δ is a very small positive number, and the unweighted output probability $q_{ij} / \sum_k q_{ik}$ similarly takes a high value $C_i = \frac{1-\delta}{K_i}$ for K_i points and a low value $D_i = \frac{\delta}{N-K_i-1}$ for other points. Denote the set where $\tilde{p}_{j|i} = A_i$ and $q_{ij} / \sum_k q_{ik} = C_i$ by $S_{TP,i}$, the set where $\tilde{p}_{j|i} = A_i$ and $q_{ij} / \sum_k q_{ik} = D_i$ by $S_{MISS,i}$, the set where $\tilde{p}_{j|i} = B_i$ and $q_{ij} / \sum_k q_{ik} = C_i$ by $S_{FP,i}$, and the set where $\tilde{p}_{j|i} = B_i$ and $q_{ij} / \sum_k q_{ik} = D_i$ by $S_{TN,i}$.

The Kullback-Leibler divergence between the conditional input distribution $\{\tilde{p}_{j|i}\}$ and the weighted conditional output distribution $\{\tilde{q}_{j|i}\}$ can then be written as

$$\begin{aligned}
&D_{\text{KL}}(\{\tilde{p}_{j|i}\}||\{\tilde{q}_{j|i}\}) && (12)\\
=&-\sum_{j \in S_{TP,i}} A_i \log \frac{d_j C_i}{G_i} - \sum_{j \in S_{MISS,i}} A_i \log \frac{d_j D_i}{G_i} && (13)
\end{aligned}$$

$$- \sum_{j\in S_{FP,i}} B_i \log \frac{d_j C_i}{G_i} - \sum_{j\in S_{TN,i}} B_i \log \frac{d_j D_i}{G_i} + \text{constant} \tag{14}$$

$$= - \sum_{j\in S_{TP,i}} \frac{1-\delta}{R_i} \log \frac{d_j(1-\delta)}{K_i G_i} \tag{15}$$

$$- \sum_{j\in S_{MISS,i}} \frac{1-\delta}{R_i} \log \frac{d_j \delta}{(N-K_i-1)G_i} \tag{16}$$

$$- \sum_{j\in S_{FP,i}} \frac{\delta}{N-R_i-1} \log \frac{d_j(1-\delta)}{K_i G_i} \tag{17}$$

$$- \sum_{j\in S_{TN,i}} \frac{\delta}{N-R_i-1} \log \frac{d_j \delta}{(N-K_i-1)G_i} + \text{constant.} \tag{18}$$

where in this simplified situation

$$G_i = \sum_k d_k \frac{q_{ik}}{\sum_l q_{il}} \tag{19}$$

$$= \sum_{k\in S_{TP,i}} d_k C_i \tag{20}$$

$$+ \sum_{k\in S_{MISS,i}} d_k D_i \tag{21}$$

$$+ \sum_{k\in S_{FP,i}} d_k C_i \tag{22}$$

$$+ \sum_{k\in S_{TN,i}} d_k D_i. \tag{23}$$

Analysis of dominating terms, and information retrieval interpretation. Assuming that each $d_j \geq 1$ (and thus also that $G_i \geq 1$), the dominating terms are the terms $\frac{1-\delta}{R_i}\log(d_j\delta) \approx \frac{1}{R_i}\log(d_j\delta)$ and the divergence simplifies to

$$D_{\text{KL}}(\{\tilde{p}_{j|i}\}||\{\tilde{q}_{j|i}\}) \tag{24}$$

$$\approx - \frac{1}{R_i} \sum_{j\in S_{MISS,i}} \log(d_j\delta) + \text{constant} \tag{25}$$

$$= \frac{N_{MISS,i}}{R_i}\left[\log\frac{1}{\delta} - \frac{\sum_{j\in S_{MISS,i}} \log d_j}{N_{MISS,i}}\right] + \text{constant} \tag{26}$$

$$= (1-\text{recall}(i)) \cdot \left[\log\frac{1}{\delta} - \frac{\sum_{j\in S_{MISS,i}} \log d_j}{N_{MISS,i}}\right] + \text{constant,} \tag{27}$$

where $N_{MISS,i} = |S_{MISS,i}|$ is the number of missed true neighbors which were not close to i on the display, and $\text{recall}(i) = 1 - \frac{N_{MISS,i}}{R_i}$ is the standard definition of recall, the proportion of true neighbors retrieved from the display out of all the original true neighbors.

The last line of (27) provides an information retrieval interpretation of the divergence between the conditional distributions. The term at left measures recall of neighbors. The term at right is a weighting term interpreted as the average cost of the misses. In the weighting term, the cost of missing high-degree neighbors is discounted, in order to allow high-degree nodes to be kept farther from each other than in the unweighted setting; this prevents high-degree nodes from crowding and allows the method to distribute the high- and low-degree nodes more evenly.

Thus, when ws-SNE minimizes the weighted average of the Kullback-Leibler divergences (27), it optimizes recall of true neighbors from the display, weighted by severity of the missed neighbors.

2.2. Analysis of the marginal divergence

First, denote the unweighted marginal probability by $q_i = \sum_j q_{ij} / \sum_{kl} q_{kl}$, and note that the weighted marginal output distribution $\tilde{q}_i$ can then be written as

$$\tilde{q}_i \tag{28}$$

$$= \frac{d_i \sum_j d_j q_{ij}}{\sum_{kl} d_k d_l q_{kl}} \tag{29}$$

$$= \frac{d_i \Big((\sum_j d_j q_{ij})(\sum_j q_{ij})^{-1} \Big) \Big((\sum_j q_{ij})(\sum_{mn} q_{mn})^{-1} \Big)}{\sum_k d_k \Big((\sum_l d_l q_{kl})(\sum_l q_{kl})^{-1} \Big) \Big((\sum_l q_{kl})(\sum_{mn} q_{mn})^{-1} \Big)} \tag{30}$$

$$= \frac{d_i q_i G_i}{\sum_k d_k q_k G_k} \;. \tag{31}$$

The weighted marginal probability attains high values for nodes i that have high degree and that are near many other nodes of high degree. Nodes that are far from the other nodes (low q_i) have low marginal weighted probability.

Now consider a simplified situation where some for some nodes the marginal probability $\tilde{p}_i$ takes a high value $A = \frac{1-\delta}{R}$ for R points and low value $B = \frac{\delta}{N-R}$ for others, where δ is a very small positive number, and on the display the unweighted marginal probability q_i takes a high value $C = \frac{1-\delta}{K}$ for K points and low value $D = \frac{\delta}{N-K}$ for others. Denote the set where $\tilde{p}_i = A$ and $q_i = C$ by S'_{TP}, the set where $\tilde{p}_i = A$ and $q_i = D$ by S'_{MISS}, the set where $\tilde{p}_i = B$ and $q_i = C$ by S'_{FP}, and the set where $\tilde{p}_i = B$ and $q_i = D$ by S'_{TN}.

The Kullback-Leibler divergence of the marginal distributions can then be written as

$$D_{\text{KL}}(\{\tilde{p}_i\}||\{\tilde{q}_i\}) \tag{32}$$

$$= - \sum_{i\in S'_{TP}} A \log \tilde{q}_i - \sum_{i\in S'_{MISS}} A \log \tilde{q}_i \tag{33}$$

$$- \sum_{i\in S'_{FP}} B\log\tilde{q}_i - \sum_{i\in S'_{TN}} B\log\tilde{q}_i + \text{constant}. \tag{34}$$

Analysis of dominating terms, and information retrieval interpretation. Assuming again that each $d_i \geq 1$ and thus also each $G_i \geq 1$, the dominating terms are the terms where $i \in S'_{MISS}$ since there the term under the logarithm is $\tilde{q}_i \propto q_i$ which is close to zero. The divergence then simplifies to

$$\begin{aligned}
&D_{\mathrm{KL}}(\{\tilde{p}_i\}||\{\tilde{q}_i\}) \quad (35)\\
\approx &- \sum_{i\in S'_{MISS}} A\log\tilde{q}_i + \text{constant} \quad (36)\\
= &- \sum_{i\in S'_{MISS}} \frac{1-\delta}{R}\log\left(\frac{\delta}{N-K}\cdot\frac{d_iG_i}{\sum_k d_kq_kG_k}\right) + \text{constant} \quad (37)
\end{aligned}$$

where on the second line we inserted the form of $\tilde{q}_i$ from (31). Further leaving out all but the dominating terms, this simplifies to

$$\begin{aligned}
&D_{\mathrm{KL}}(\{\tilde{p}_i\}||\{\tilde{q}_i\}) \quad (38)\\
\approx &- \frac{1}{R}\sum_{i\in S'_{MISS}} \log\left(\delta\cdot\frac{d_iG_i}{\sum_k d_kq_kG_k}\right) + \text{constant} \quad (39)\\
= &\frac{N_{MISS}}{R}\left[\log\frac{1}{\delta} - \frac{1}{N_{MISS}}\sum_{i\in S'_{MISS}}\log\left(\frac{d_iG_i}{\sum_k d_kq_kG_k}\right)\right]\\
&+ \text{constant} \quad (40)\\
= &(1-\text{recall})\left[\log\frac{1}{\delta} - \frac{1}{N_{MISS}}\sum_{i\in S'_{MISS}}\log\left(\frac{d_iG_i}{\sum_k d_kq_kG_k}\right)\right]\\
&+ \text{constant} \quad (41)
\end{aligned}$$

where $N_{MISS} = |S'_{MISS}|$ is the number of initial nodes not chosen based on the display that were chosen based on original data, and recall $= 1 - \frac{N_{MISS}}{R}$ is the corresponding standard definition of recall.

The last line of (41) provides an information retrieval interpretation of the divergence between the marginal distributions. The term at left evaluates recall of the initial nodes. The term on the right is again a weighting term evaluating the cost of initial nodes left out (missed). The misses here denote points with low q_i that are far from other nodes and would not be chosen as initial nodes based on the visual arrangement on the display. The weighting term discounts missing high-degree nodes (more precisely, nodes with high degree d_i and high degree G_i of nearby neighbors), and thus allows them to be placed in sparser areas of the display than in an unweighted setting; this allows the high and low-degree nodes to be placed more evenly on the display.

Conclusion. Based on the separation of the ws-SNE cost function into two kinds of divergences, and the information retrieval interpretation of both kinds of divergences, we conclude that the ws-SNE cost function corresponds to a two-stage information retrieval task of first choosing initial nodes and then choosing neighbors for them. For both retrieval tasks, the method minimizes cost of errors in such retrieval which are dominated by costs of misses, and the degree (importance) of nodes is taken into account by discounting the costs of high-degree nodes in order to avoid crowding of the high-degree nodes on the display.

3. Additional visualizations

Figure 1 in the paper is high resolution, and readers can find more details by zooming in. The visualizations in this supplemental document provide extra information that cannot fit to the paper due to space limit. Moreover, we also provide the results of two other datasets.

- Figures 1 to 4 shows the large visualization of the `worldtrade` dataset using ws-SNE. It also displays the countries names and different node sizes to facilitate comparison to our common knowledge.
- Figures 5 to 8 shows the large visualization of the `usair97` dataset using ws-SNE. It also displays the airport names and different node sizes to facilitate comparison to our common knowledge.
- Figures 9 to 12 shows the large visualization of the `mirex07` dataset using ws-SNE. It also displays the class names and different node sizes to facilitate comparison to our common knowledge.
- Figure 13 shows the geographical layout (ground truth) of the `luxembourg` dataset, which can be used to compared the sixth row in Figure 1 in the paper.
- Figure 14 shows the visualizations of the `jazz` dataset using graphviz, LinLog, t-SNE, and ws-SNE.
- Figure 15 shows the visualizations of the `ca-GrQc` dataset using graphviz, LinLog, t-SNE, and ws-SNE.
- Figure 16 shows the full visualization of the `shuttle` dataset using EE with $\lambda = 1$.
- Figure 17 shows the full visualization of the `shuttle` dataset using ws-SNE.
- Figure 18 shows the full visualization of the `MNIST` dataset using EE with $\lambda = 1$.

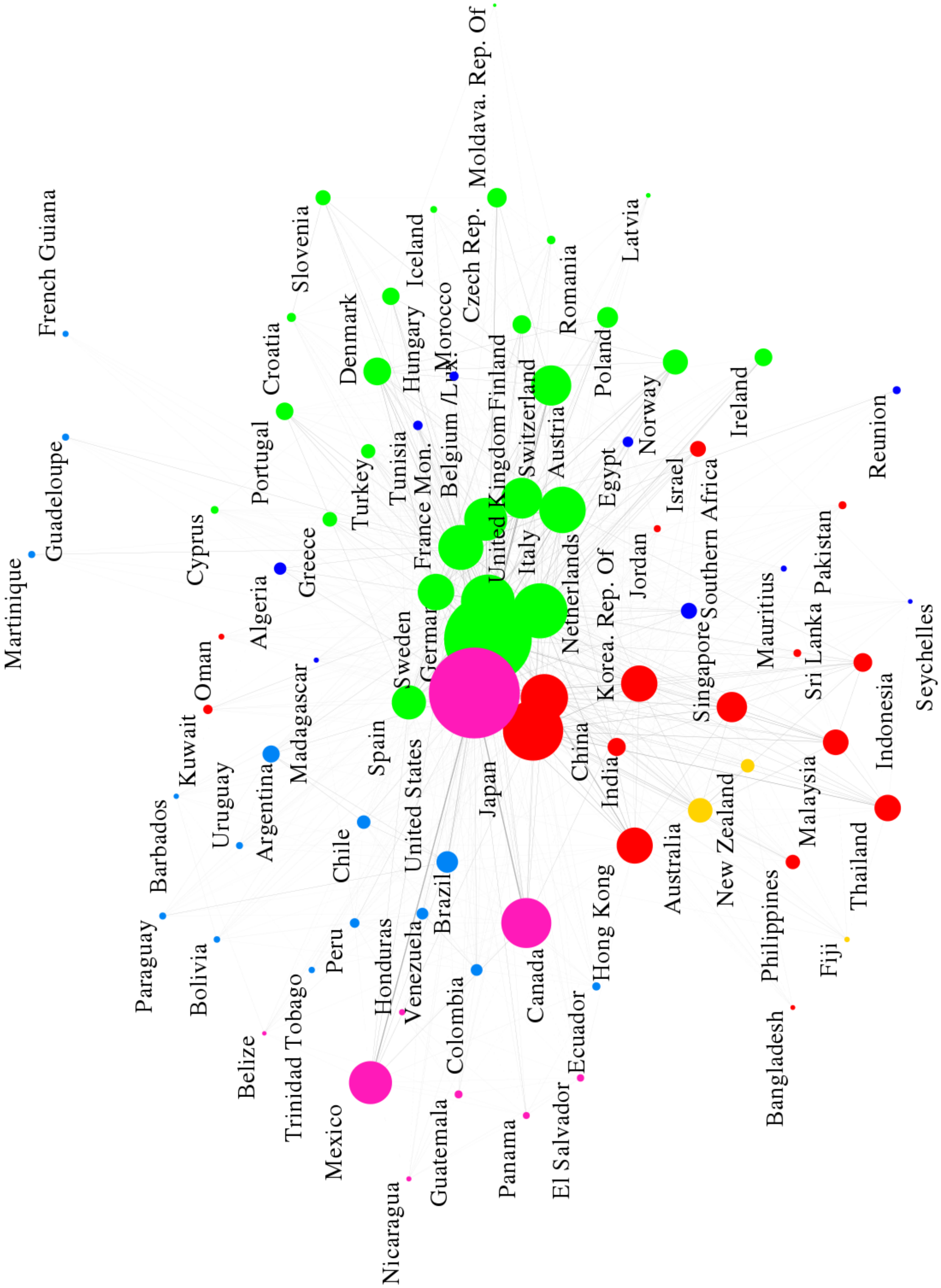

Figure 1. Visualizations with text labels for the `worldtrade` dataset using graphviz. The node size is proportional to the square root of its degree.

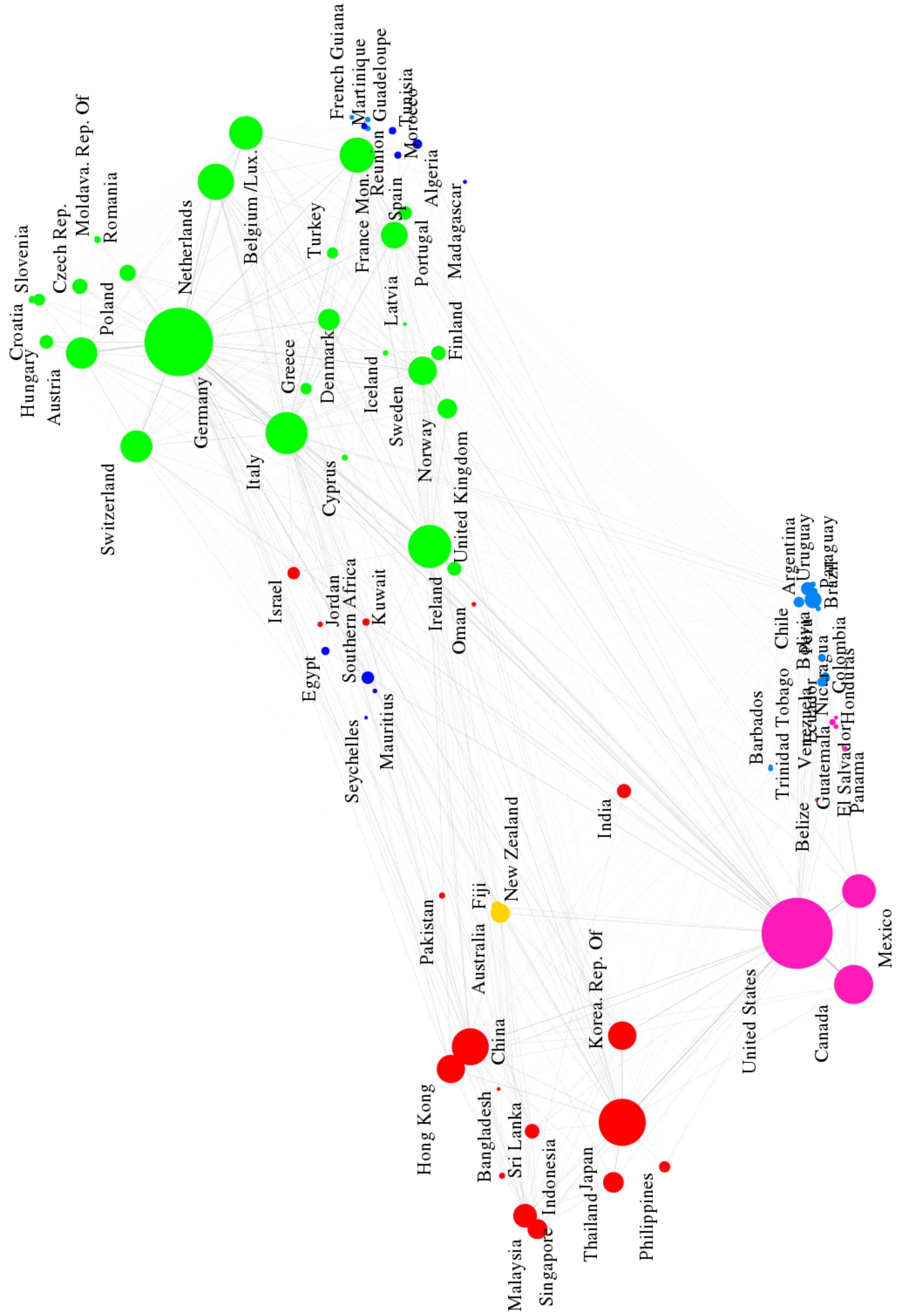

Figure 2. Visualizations with text labels for the `worldtrade` dataset using LinLog. The node size is proportional to the square root of its degree.

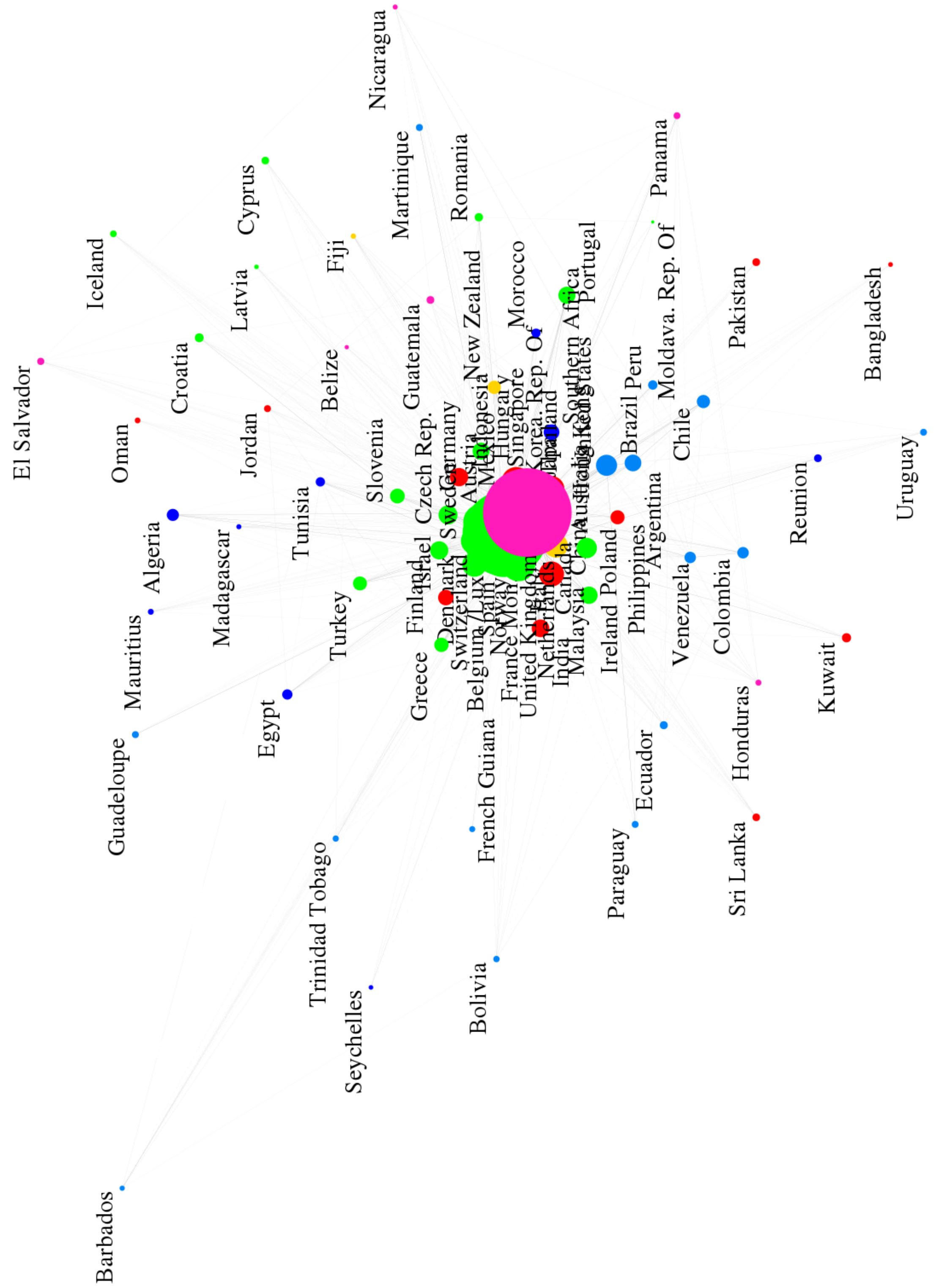

Figure 3. Visualizations with text labels for the `worldtrade` dataset using t-SNE. The node size is proportional to the square root of its degree.

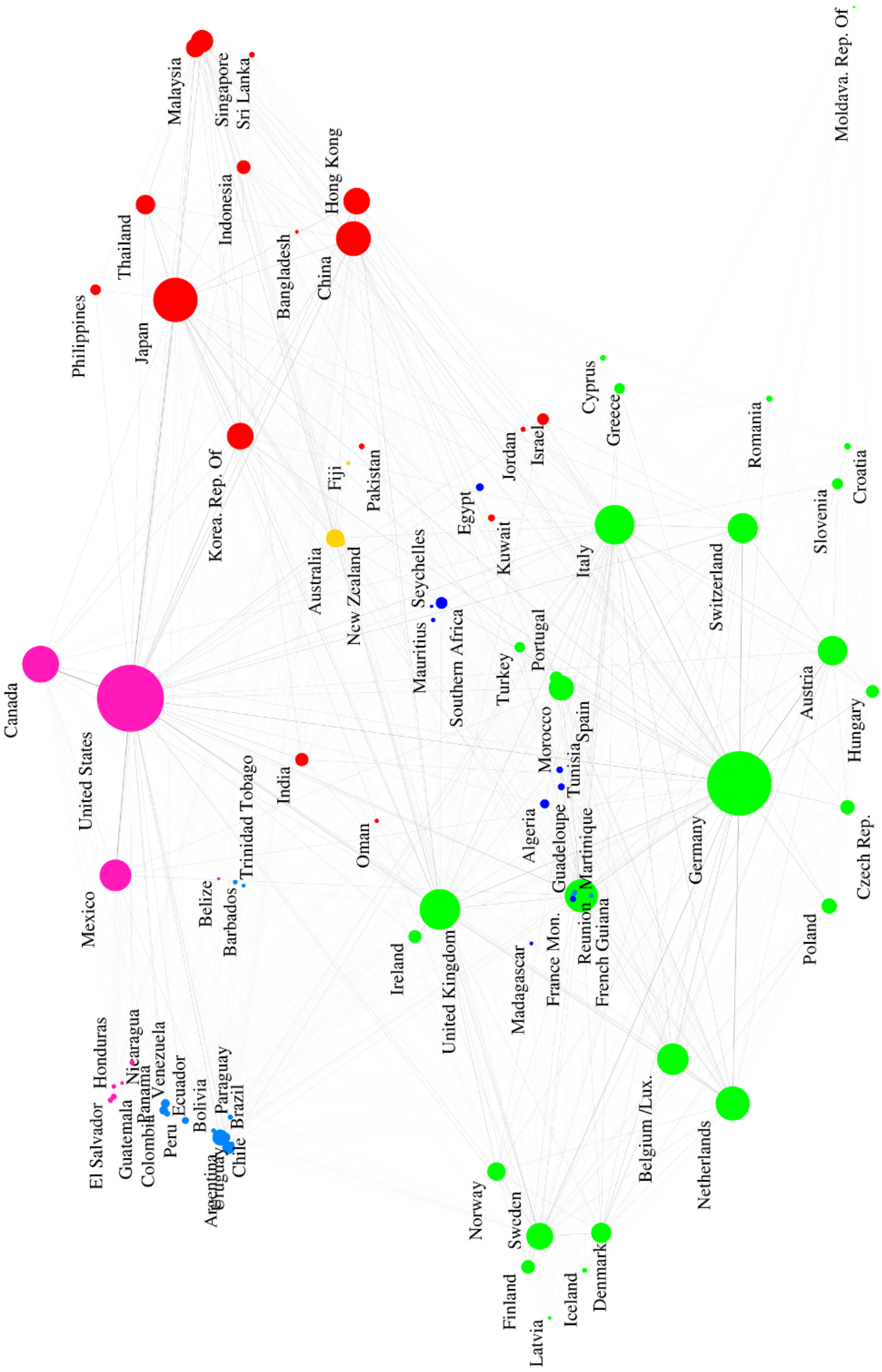

Figure 4. Visualizations with text labels for the `worldtrade` dataset using ws-SNE (Cauchy kernel). The node size is proportional to the square root of its degree.

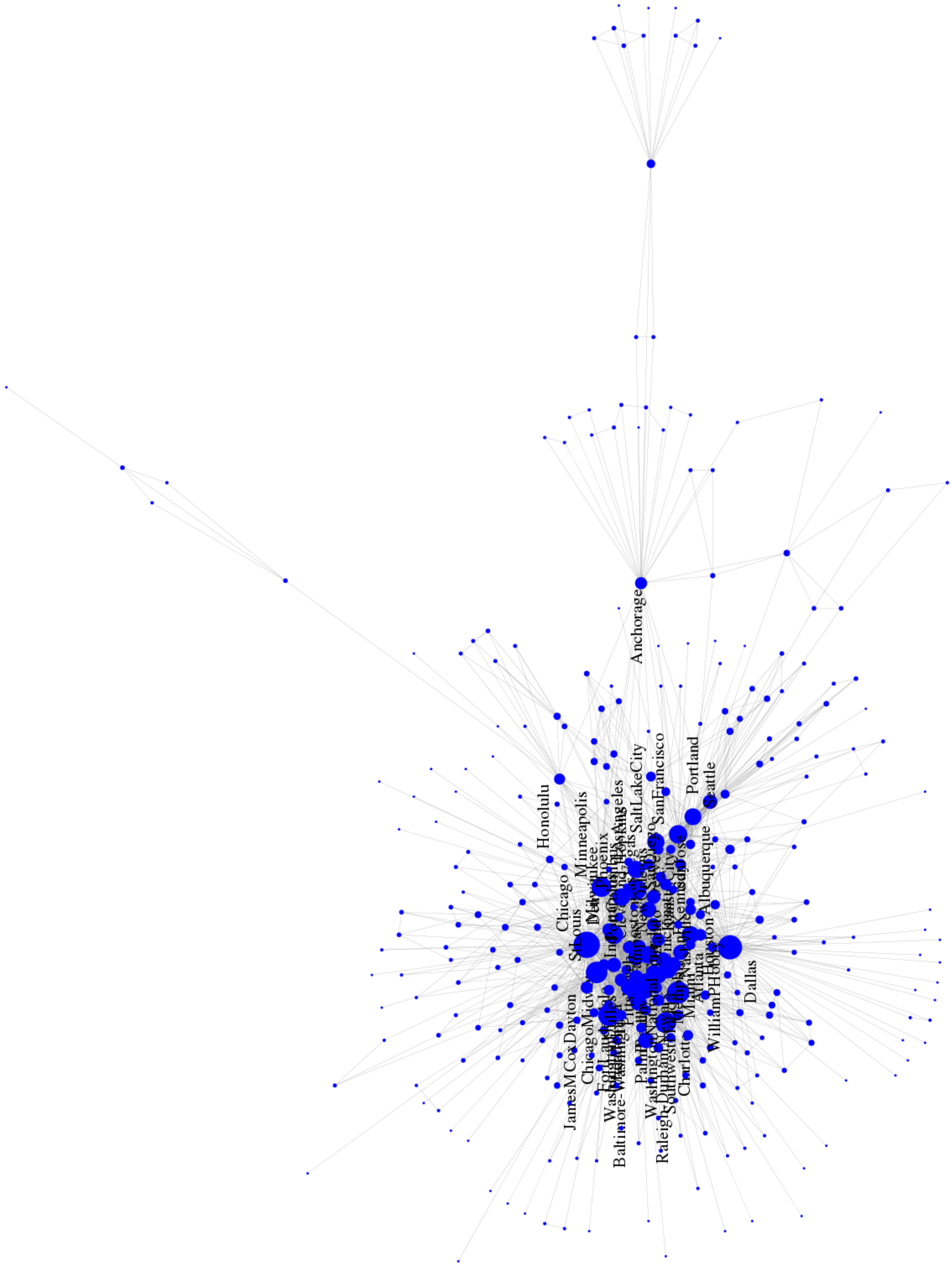

Figure 5. Visualizations with text labels for the `usair97` dataset using graphviz. The node size is proportional to the square root of its degree.

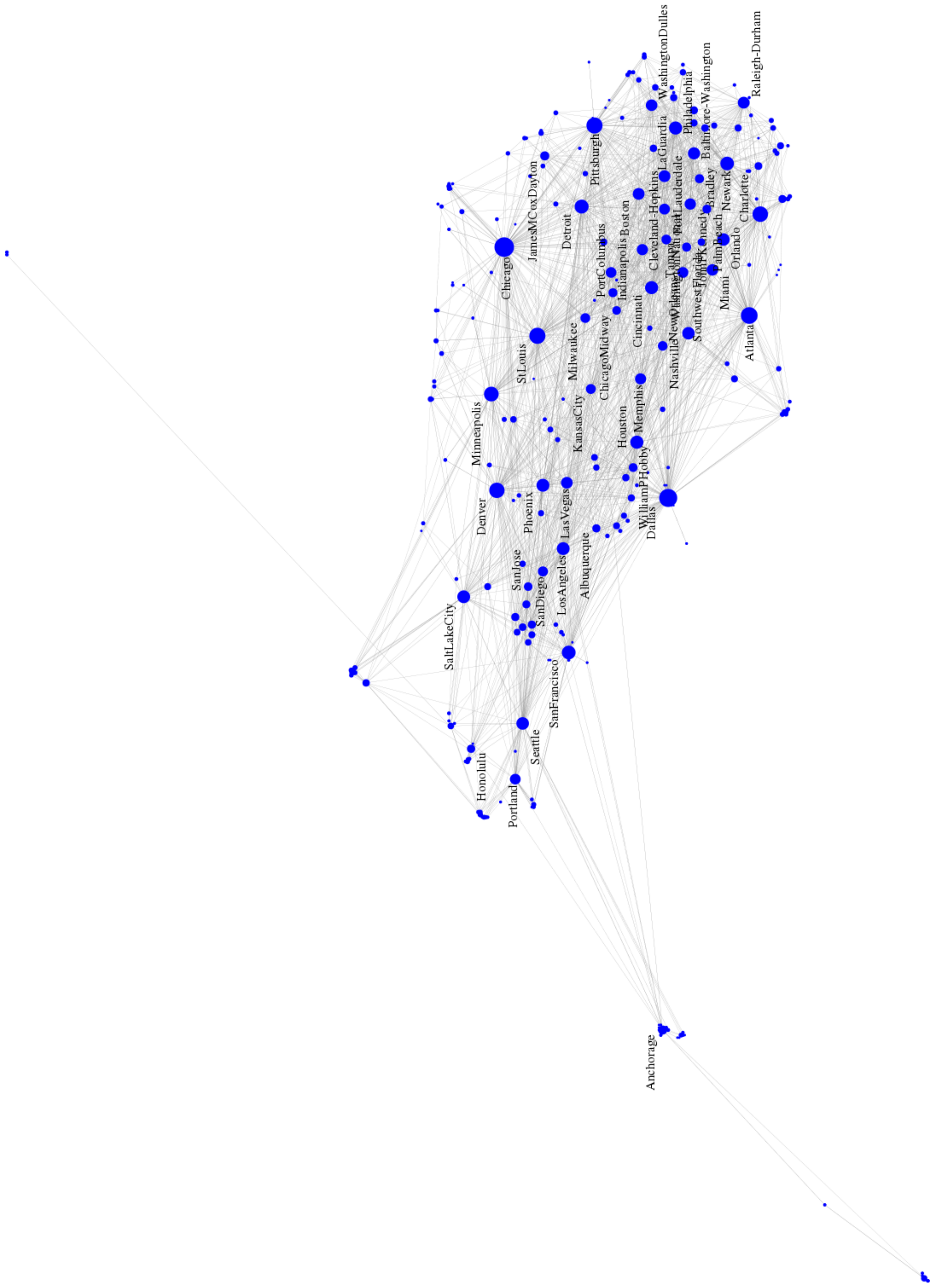

Figure 6. Visualizations with text labels for the `usair97` dataset using LinLog. The node size is proportional to the square root of its degree.

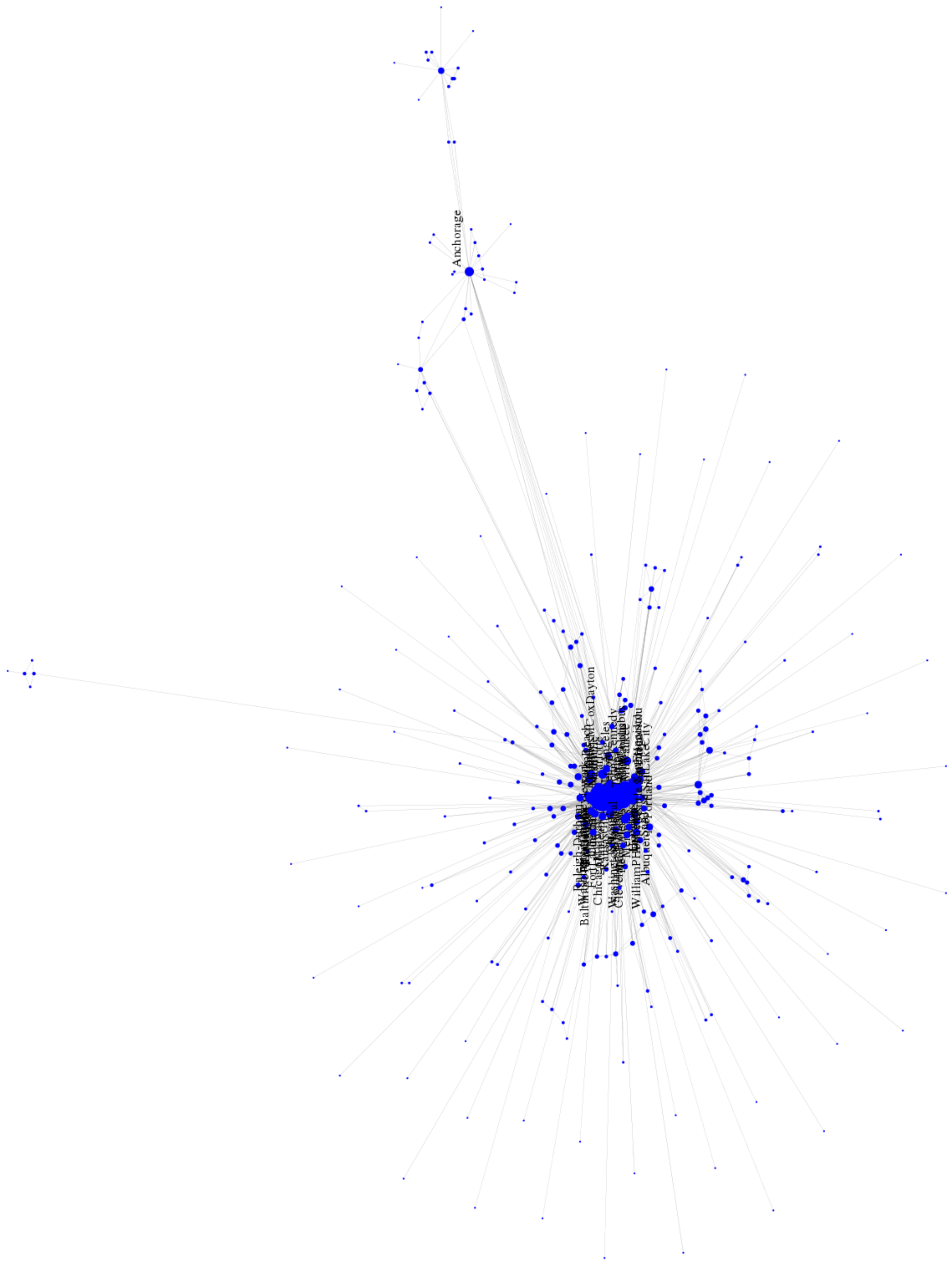

Figure 7. Visualizations with text labels for the `usair97` dataset using t-SNE. The node size is proportional to the square root of its degree.

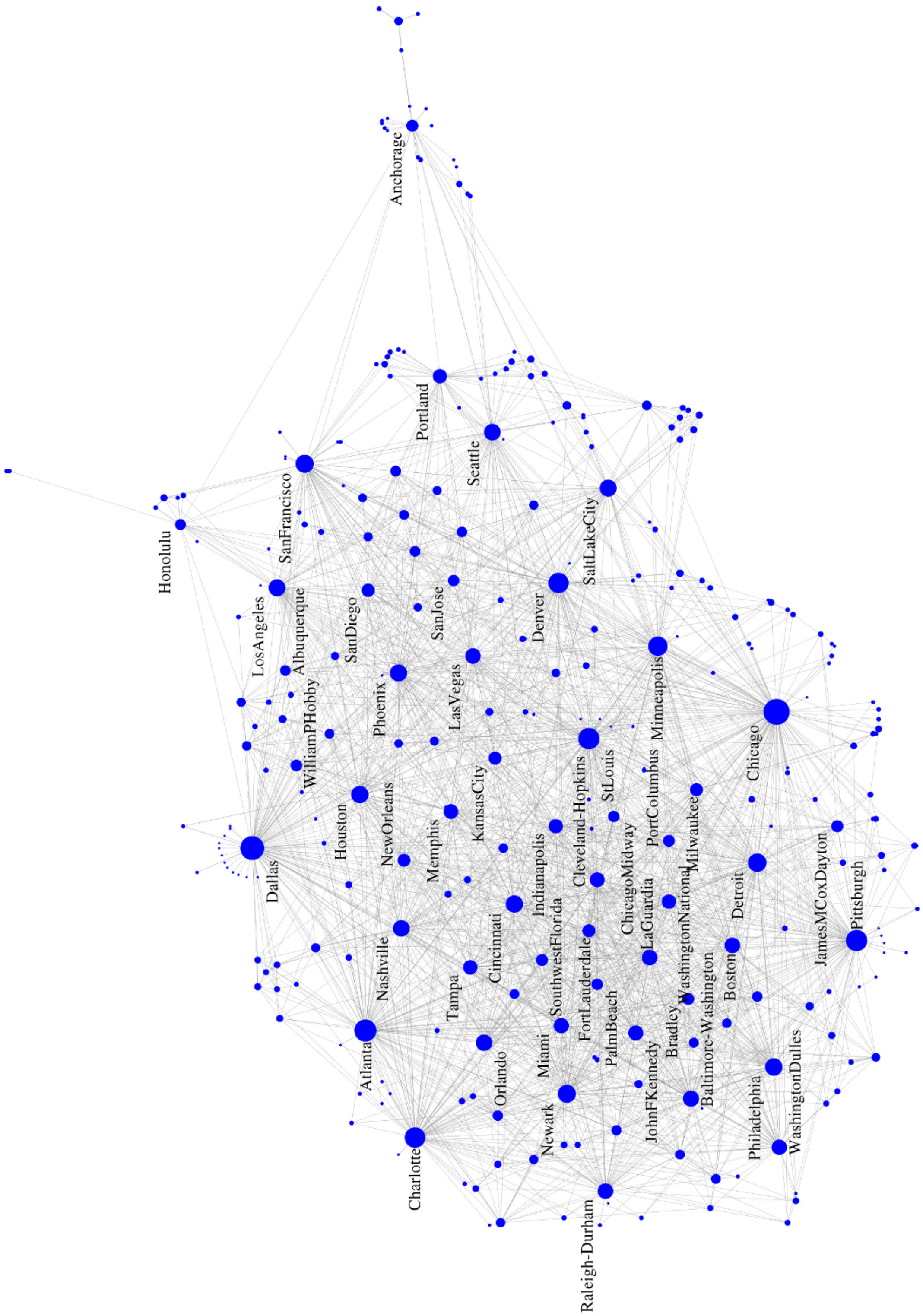

Figure 8. Visualizations with text labels for the `usair97` dataset using ws-SNE (Cauchy kernel). The node size is proportional to the square root of its degree.

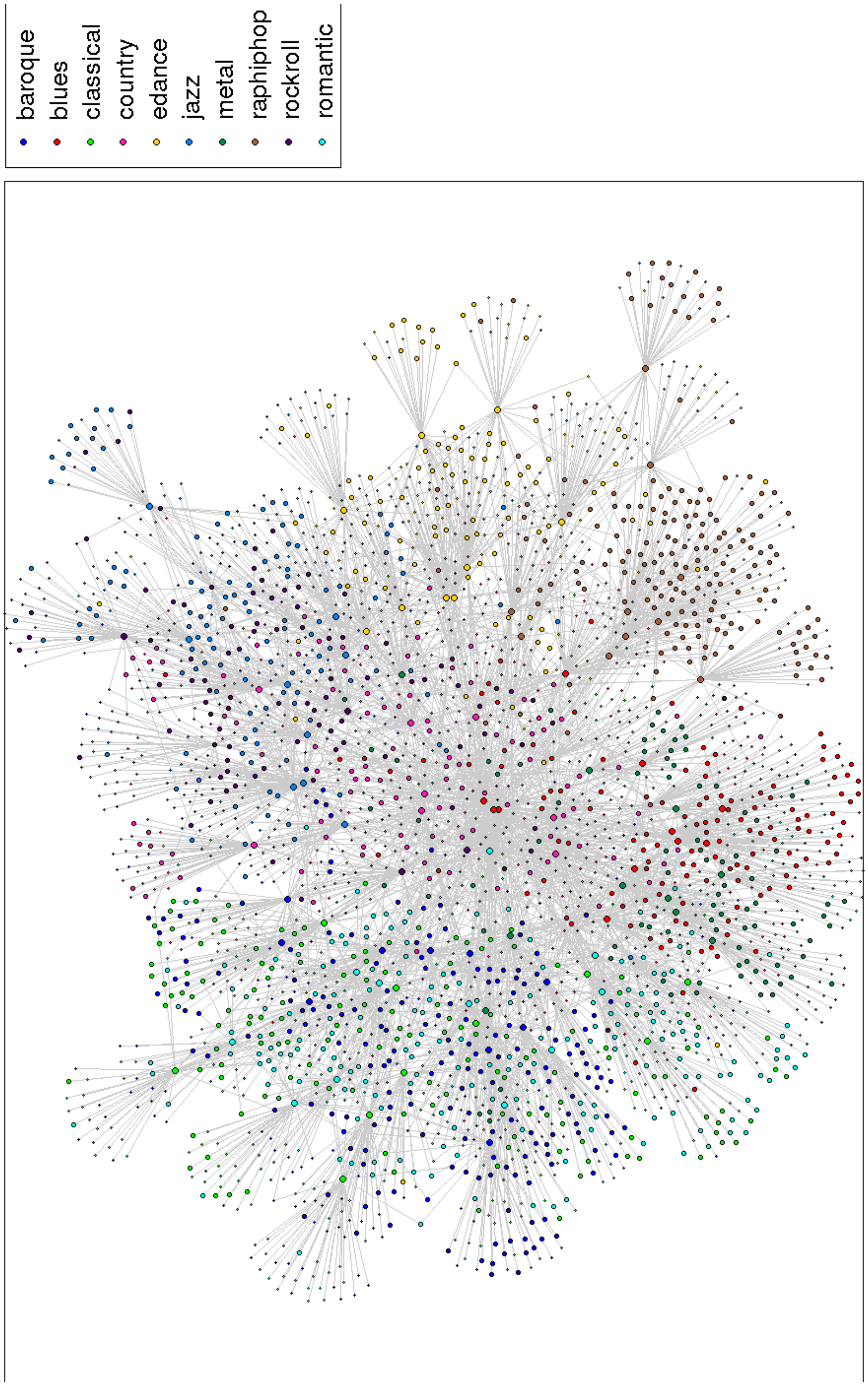

Figure 9. Visualizations with class names (shown by legend) for the `mirex07` dataset using graphviz. The node size is proportional to the square root of its degree.

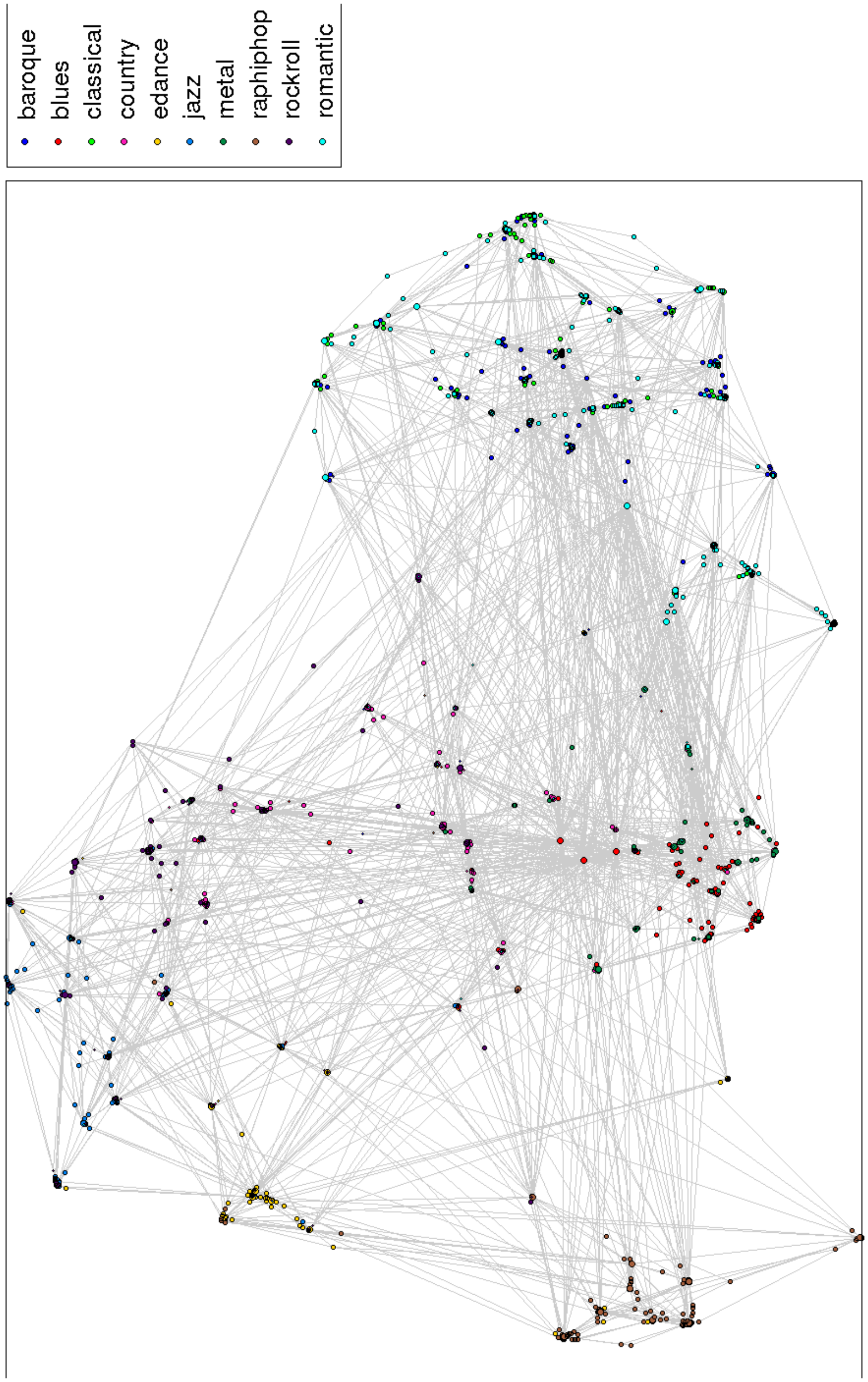

Figure 10. Visualizations with class names (shown by legend) for the `mirex07` dataset using LinLog. The node size is proportional to the square root of its degree.

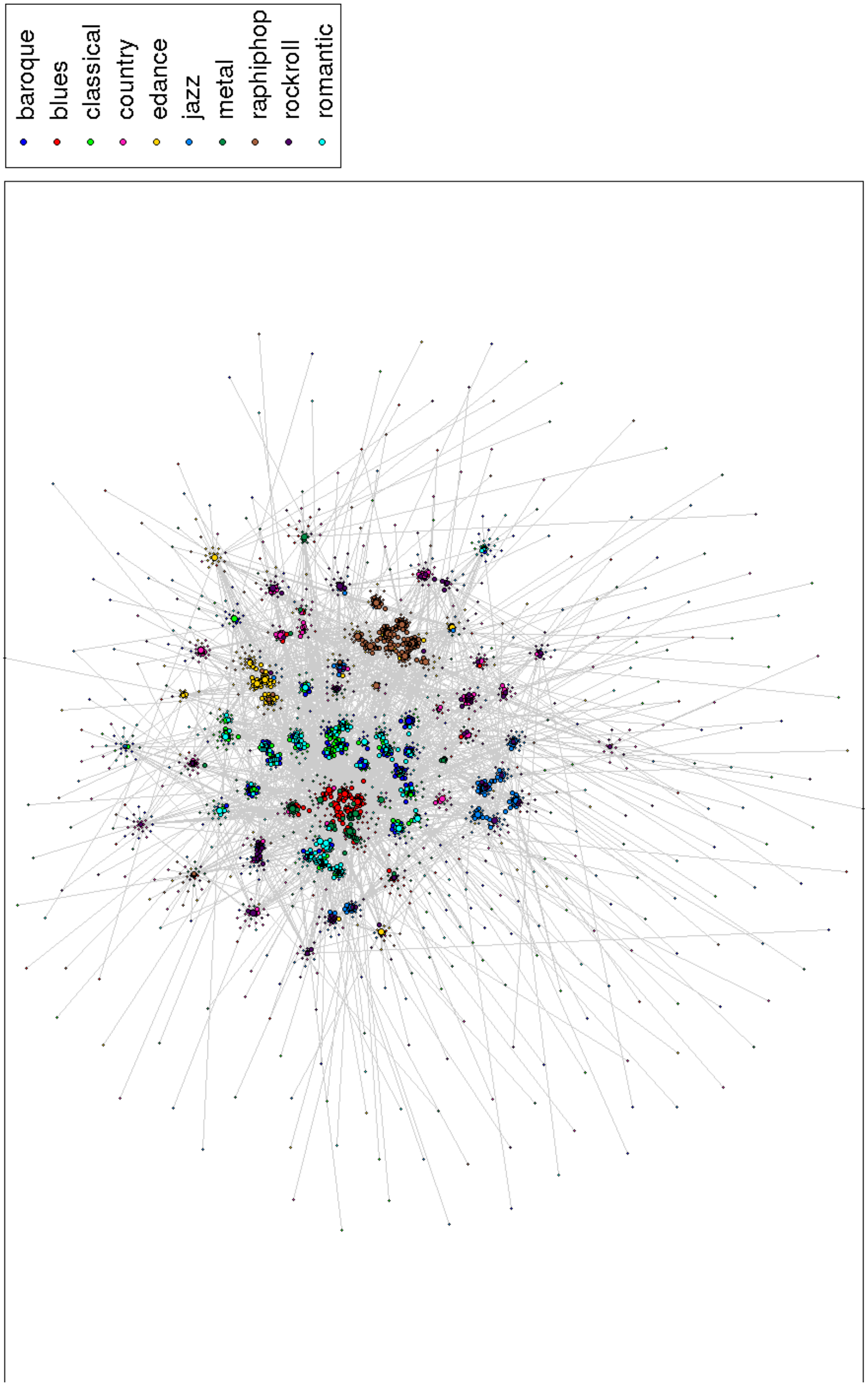

Figure 11. Visualizations with class names (shown by legend) for the `mirex07` dataset using t-SNE. The node size is proportional to the square root of its degree.

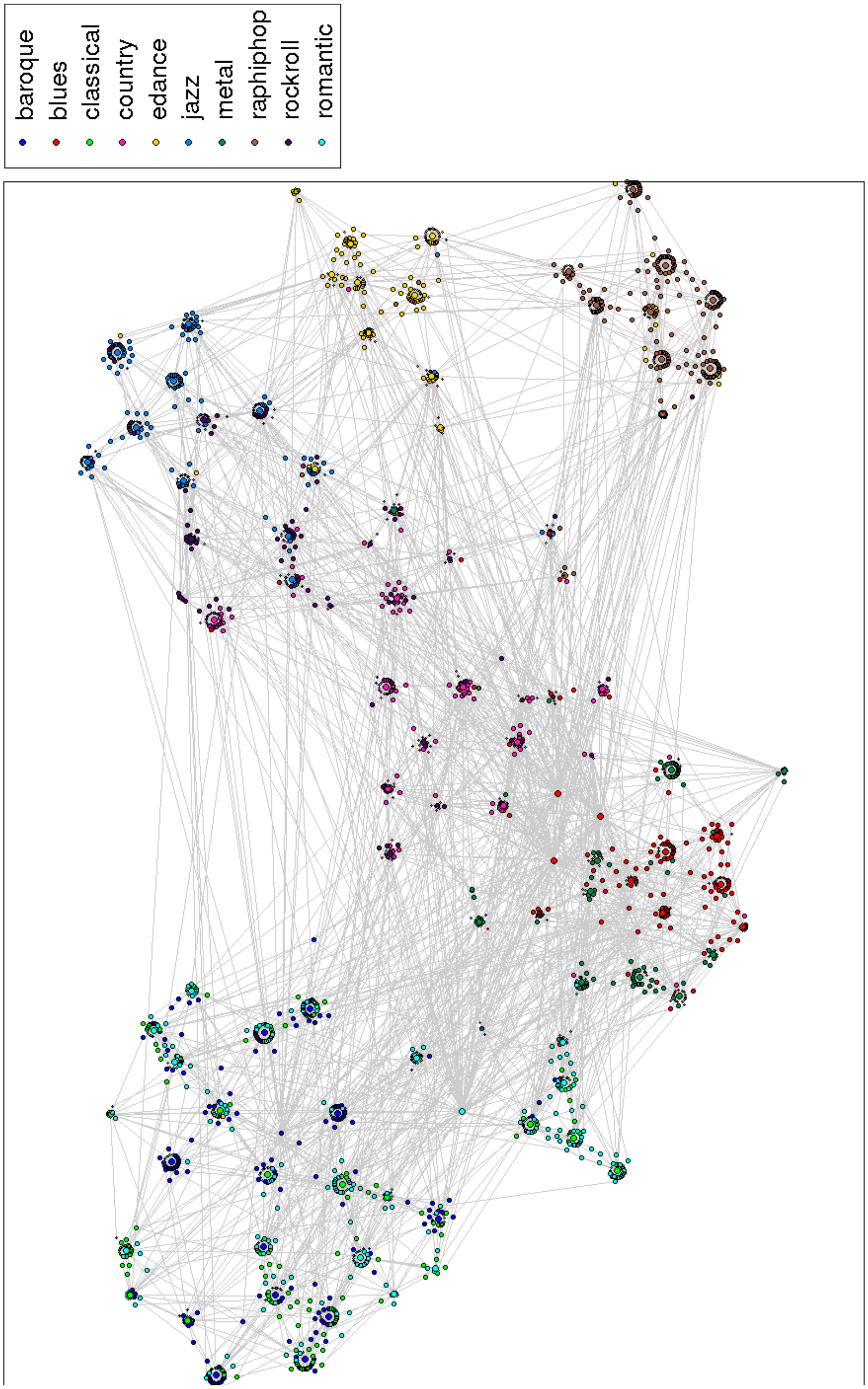

Figure 12. Visualizations with class names (shown by legend) for the `mirex07` dataset using ws-SNE (Cauchy kernel). The node size is proportional to the square root of its degree.

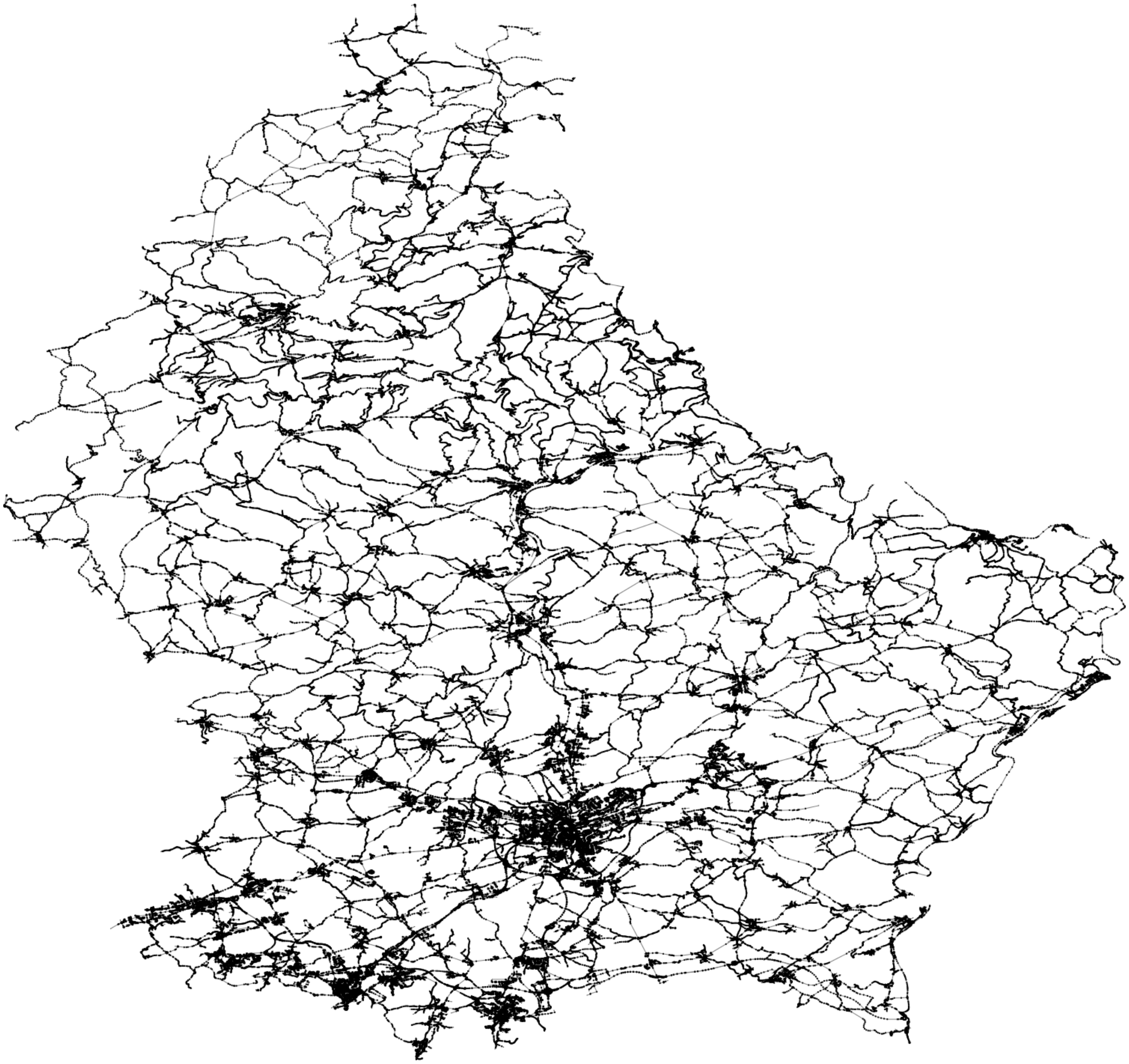

Figure 13. Geographical layout (ground truth) of the `luxembourg` dataset.

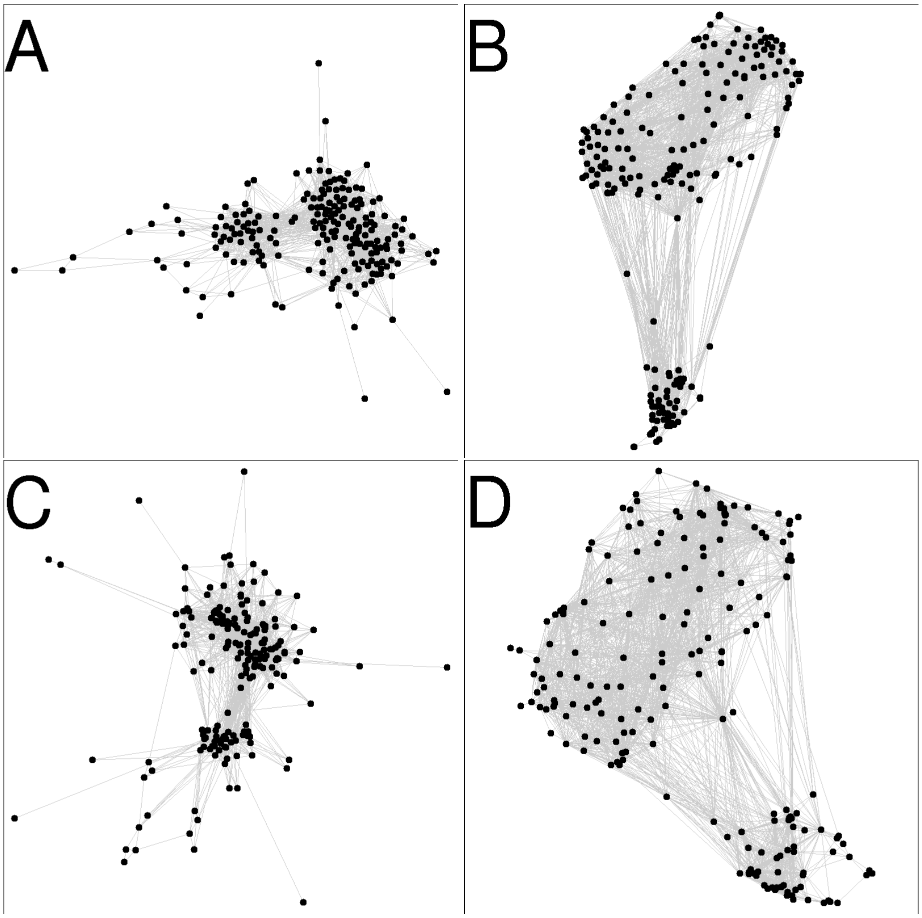

Figure 14. Visualizations of the `jazz` dataset using A) graphviz, B) LinLog, C) t-SNE, and D) ws-SNE. Cauchy kernel was used in ws-SNE. A desired layout should clearly reveal the two musician groups, LinLog and ws-SNE are able to do so.

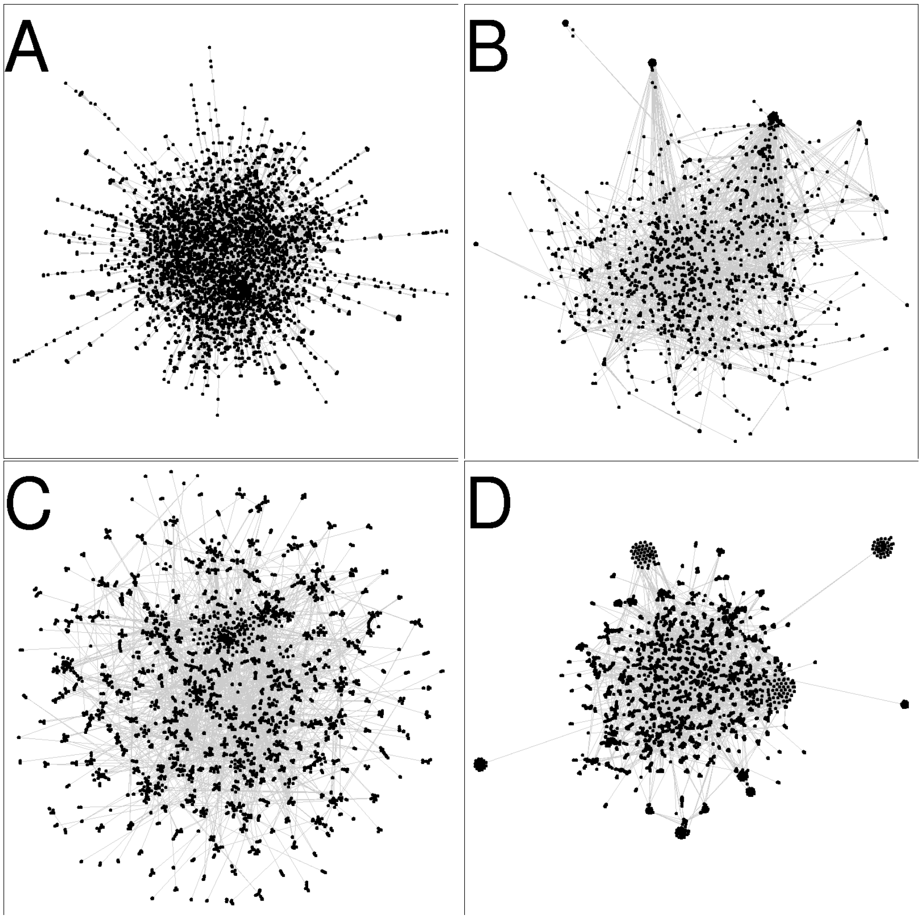

Figure 15. Visualizations of the `ca-GrQc` dataset using A) graphviz, B) LinLog, C) t-SNE, and D) ws-SNE. Cauchy kernel was used in ws-SNE. Leskovec et al. (2009) has shown that for large coauthor networks, there is a central big community, as well as one or more small communities in periphery. Thus we expect that a good layout will here as well show both a central community and smaller peripheral communities as ws-SNE does.

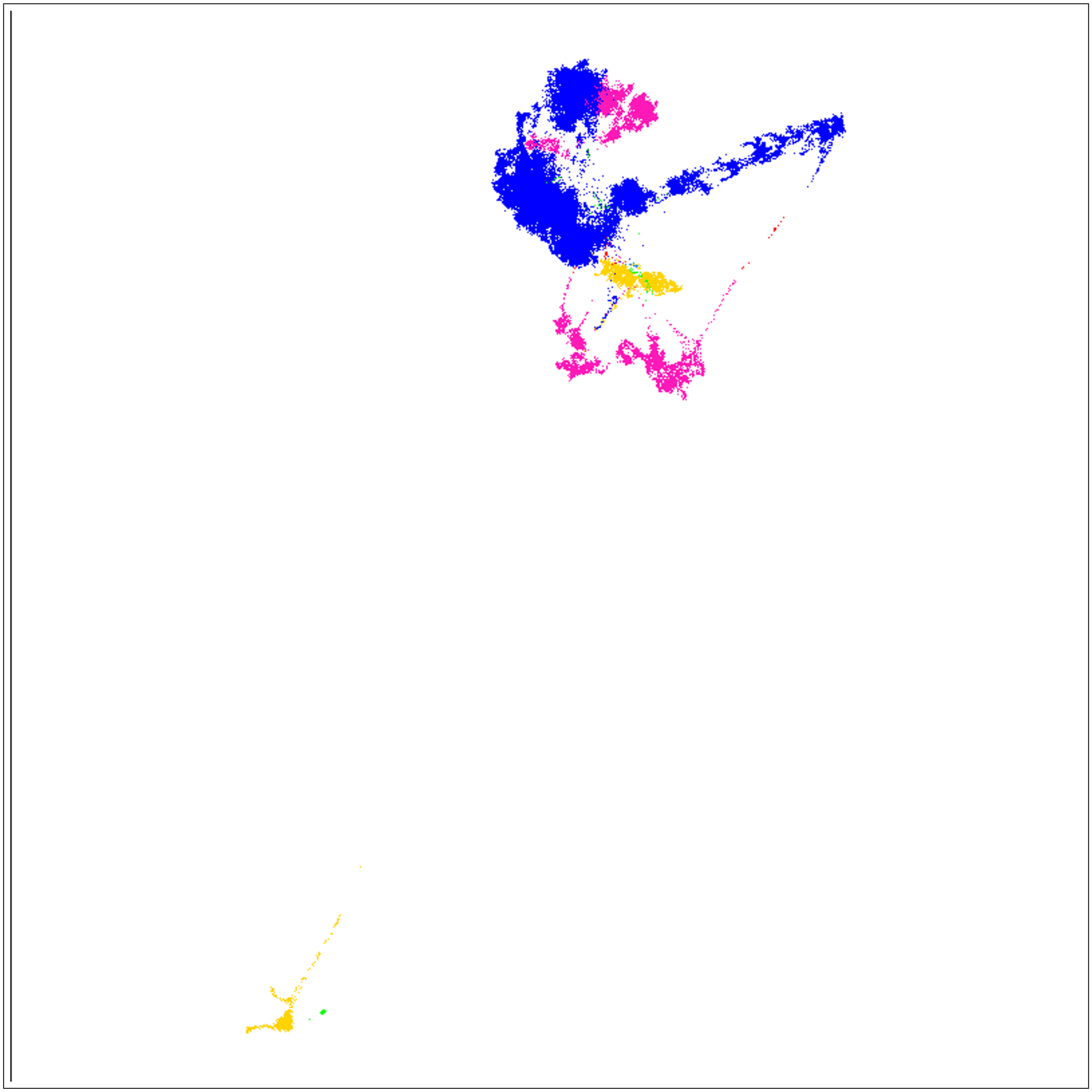

Figure 16. Full visualization of `shuttle` using EE with $\lambda = 1$.

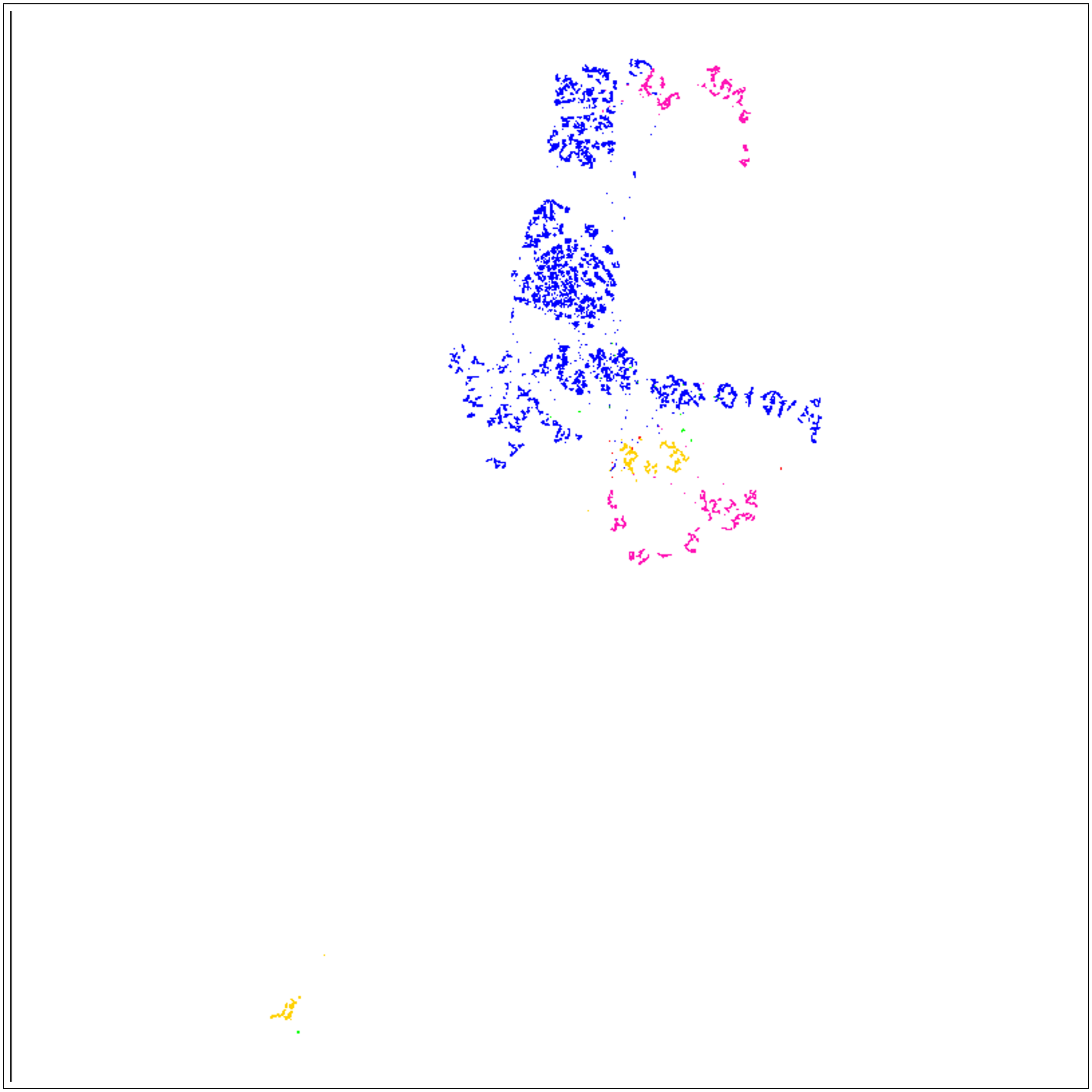

Figure 17. Full visualization of `shuttle` using ws-SNE with the Cauchy kernel.

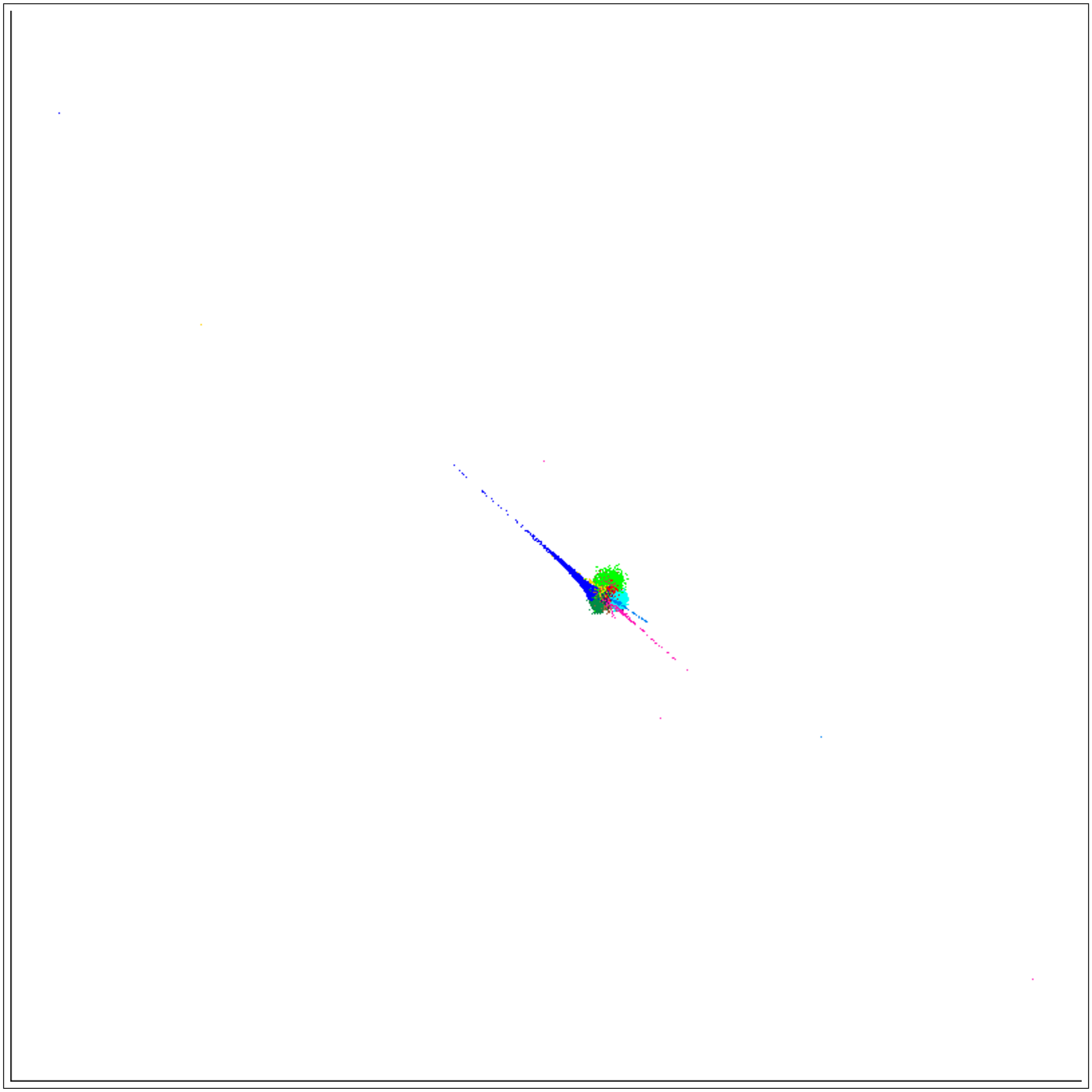

Figure 18. Full visualization of MNIST using EE with $\lambda = 1$.

References

Chen, Y., Garcia, E., Gupta, M., Rahimi, A., and Cazzanti, L. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

Mallat, S. Group invariant scattering. *Communications in Pure and Applied Mathematics*, 2012.