Equivalence between LINE and Matrix Factorization

Qiao Wang, Zheng Wang,* Xiaojun Ye
School of Software, Tsinghua University, Beijing, China
{wangqiao15, zheng-wang13}@mails.tsinghua.edu.cn
{yexj}@tsinghua.edu.cn

November 9, 2017

Abstract

LINE [1], as an efficient network embedding method, has shown its effectiveness in dealing with large-scale undirected, directed, and/or weighted networks. Particularly, it proposes to preserve both the local structure (represented by First-order Proximity) and global structure (represented by Second-order Proximity) of the network. In this study, we prove that LINE with these two proximities (LINE(1st) and LINE(2nd)) are actually factoring two different matrices separately. Specifically, LINE(1st) is factoring a matrix $M^{(1)}$, whose entries are the doubled Pointwise Mutual Information (PMI) of vertex pairs in undirected networks, shifted by a constant. LINE(2nd) is factoring a matrix $M^{(2)}$, whose entries are the PMI of vertex and context pairs in directed networks, shifted by a constant. We hope this finding would provide a basis for further extensions and generalizations of LINE.

1 Notation and Definition

Given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each edge $e \in \mathcal{E}$ is an ordered pair $e = (v_i, v_j)$ and has an associated weight $w_{ij} > 0$. In directed networks, the in-degree and out-degree of vertex v_i are denoted as $deg^-(v_i)$ and $deg^+(v_i)$ respectively. In addition, in undirected networks, the degree of vertex v_i is denoted as $deg(v_i)$.

The first-order proximity [1] characterizes the local structure similarity between vertices. More specifically, if $(v_i, v_j) \in \mathcal{E}$, w_{ij} indicates the first-order proximity between v_i and v_j, otherwise their first-order proximity is 0.

The second-order proximity [1] characterizes the global structure similarity between vertices. Mathematically, let $p_i = (w_{i1}, ..., w_{i|\mathcal{V}|})$ represent the first-order proximity between v_i and the other vertices, then the second-order proximity between v_i and v_j is characterized by the similarity between p_i and p_j.

To simultaneously preserve these two proximities, Tang et al. [1] train the LINE model which preserves the first-order proximity (denoted as LINE(1st))

---

*Corresponding author: Zheng Wang.

and second-order proximity (denoted as LINE(2nd)) separately and then concatenate the embeddings learned by these two methods.

2 Proof

Levy and Goldberg [2] have shown that Skip-Gram with Negative Sampling [3] is implicitly factoring a word-context matrix. Similarly, in this study we prove that LINE(1st) and LINE(2nd) are actually factoring two different matrices separately. For ease of understanding, we first give the proof of LINE(2nd) and then give the proof of LINE(1st).

2.1 Equivalence of LINE(2nd) and Matrix Factorization

LINE(2nd) assumes the given network is directed (an undirected edge can be treated as two directed edges with opposite directions and equal weights), and for each directed edge (v_i, v_j) it defines v_j as the "context" of v_i. As such, on one hand each vertex $v_i \in \mathcal{V}$ is embedded into a d-dimensional vector $\overrightarrow{v_i}$ ($d \ll |\mathcal{V}|$) when it plays as the vertex itself; on the other hand it is embedded into a d-dimensional vector $\overrightarrow{u_i}$ when it plays as a "context" of the others. Let V denote a $d \times |\mathcal{V}|$ matrix whose i-th column is the vertex embedding $\overrightarrow{v_i}$ and U denote a $d \times |\mathcal{V}|$ matrix whose j-th column is the "context" embedding $\overrightarrow{u_j}$. We will figure out that LINE(2nd) is factoring a matrix $M^{(2)} = V^T U$.

According to [1], LINE(2nd) minimizes the following objective function:

$$O_2 = - \sum_{(i,j)\in\mathcal{E}} w_{ij} \log p_2(v_j|v_i) = - \sum_{(i,j)\in\mathcal{E}} w_{ij} \log \frac{exp(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i})}{\sum_{k=1}^{|\mathcal{V}|} exp(\overrightarrow{u_k}^T \cdot \overrightarrow{v_i})} \quad (1)$$

Optimizing Eq. 1 is time-consuming, since it requires the summation over all vertices when calculating the conditional probability $p_2(\cdot|v_i)$. Therefore, LINE(2nd) adopts the negative sampling approach [3], which replaces each $\log p_2(v_j|v_i)$ in Eq. 1 with the following objective function:

$$\log \sigma(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) + k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i}), \quad (2)$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function, $P_n(v)$ is a negative sampling distribution, and k defines the number of negative edges.

Next, by substituting Eq. 2 into Eq. 1, we can rewrite the objective function of LINE(2nd) as:

$$O_2 = - \sum_{(i,j)\in\mathcal{E}} w_{ij} [\log \sigma(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) + k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i})]$$

$$= - \sum_{(i,j)\in\mathcal{E}} w_{ij} \log \sigma(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) - \sum_{(i,j)\in\mathcal{E}} w_{ij} \cdot k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i})$$

$$= - \sum_{(i,j)\in\mathcal{E}} w_{ij} \log \sigma(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) - \sum_{v_i \in \mathcal{V}} deg^+(v_i) \cdot k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i})$$

$$(3)$$

To simplify the analysis, here we set the negative sampling distribution $P_n(v) \propto deg^-(v_n)$ [1]. Hence, the expectation term in Eq. 3 can be specified as follows:

$$E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i}) = \sum_{v_n \in V} \frac{deg^-(v_n)}{\sum_{v_n \in \mathcal{V}} deg^-(v_n)} \cdot \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i})$$

$$= \frac{deg^-(v_j)}{\sum_{v_n \in \mathcal{V}} deg^-(v_n)} \cdot \log \sigma(-\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) + \sum_{v_n \in \mathcal{V} \setminus \{v_j\}} \frac{deg^-(v_n)}{\sum_{v_n \in \mathcal{V}} deg^-(v_n)} \cdot \log \sigma(-\overrightarrow{u_n}^T \cdot \overrightarrow{v_i})$$

$$(4)$$

As each product $\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}$ is independent with the others, we can gain the local objective for a specific (v_i, v_j) pair by combining Eqs. 3 and 4:

$$\ell(v_i, v_j) = w_{ij} \cdot \log \sigma(\overrightarrow{u_j}^T \cdot \overrightarrow{v_i}) + deg^+(v_i) \cdot k \cdot \frac{deg^-(v_j)}{\sum_{v_n \in \mathcal{V}} deg^-(v_n)} \cdot \log \sigma(-\overrightarrow{u_j}^T \cdot \overrightarrow{v_i})$$

$$(5)$$

To minimize the objective function of LINE(2nd) (i.e., Eq. 1), we must maximize $\ell(v_i, v_j)$. As such, we define $x = \overrightarrow{u_j}^T \cdot \overrightarrow{v_i}$ and get the derivative of $\ell(v_i, v_j)$ with respect to x:

$$\frac{\partial \ell}{\partial x} = w_{ij} \cdot \sigma(-x) - deg^+(v_i) \cdot k \cdot \frac{deg^-(v_j)}{\sum_{v_n \in \mathcal{V}} deg^-(v_n)} \cdot \sigma(x) \qquad (6)$$

Comparing the derivative to zero, we have:

$$x = \overrightarrow{u_j}^T \cdot \overrightarrow{v_i} = \log \frac{w_{ij} \cdot \sum_{v_n \in \mathcal{V}} deg^-(v_n)}{deg^+(v_i) \cdot deg^-(v_j)} - \log k \qquad (7)$$

Notably, the expression $\log \frac{w_{ij} \cdot \sum_{v_n \in \mathcal{V}} deg^-(v_n)}{deg^+(v_i) \cdot deg^-(v_j)}$ is the Pointwise Mutual Information (PMI) [4] of vertex pair (v_i, v_j) in directed networks.

Overall, therefore, we can characterize the matrix $M^{(2)}$ that LINE(2nd) is actually factoring:

$$M_{ij}^{(2)} = \overrightarrow{u_j}^T \cdot \overrightarrow{v_i} = PMI(v_i, v_j) - \log k \qquad (8)$$

2.2 Equivalence of LINE(1st) and Matrix Factorization

LINE(1st) is only applicable for undirected networks. Each vertex $v_i \in \mathcal{V}$ is embedded into a d-dimensional vector $\overrightarrow{v_i}$ ($d \ll |\mathcal{V}|$) in this method. Let V denote a $d \times |\mathcal{V}|$ matrix whose i-th column is $\overrightarrow{v_i}$. We will figure out that LINE(1st) is factoring a matrix $M^{(1)} = V^T V$.

According to [1], LINE(1st) minimizes the following objective function:

$$O_1 = -\sum_{(i,j) \in \mathcal{E}} w_{ij} \log p_1(v_i, v_j) = -\sum_{(i,j) \in \mathcal{E}} w_{ij} \log \frac{1}{1 + exp(-\overrightarrow{v_i}^T \cdot \overrightarrow{v_j})} \qquad (9)$$

---

[1]LINE(2nd) sets the negative sampling distribution $P_n(v) \propto deg^+(v_n)^{3/4}$. In our proof, we replace it with $P_n(v) \propto deg^-(v_n)$. On one hand, for simplicity, we use the unigram distribution instead of its 3/4 power following [2]. On the other hand, according to the definition of negative sampling [3], the sampled negative edges for (v_i, v_j) should have the same starting point (i.e., v_i). Therefore, in this proof, we draw negative samples according to the in-degree of vertices.

To avoid the trivial solution, LINE(1st) also uses the negative sampling approach (specified in Eq. 2) by just replacing $\overrightarrow{u_j}$ with $\overrightarrow{v_j}$. More specifically, LINE(1st) replaces each $\log p_1(v_i, v_j)$ in Eq. 9 with the following objective function:

$$\log \sigma(\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) + k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i}), \tag{10}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function, $P_n(v)$ is a negative sampling distribution, and k defines the number of negative edges.

Next, by substituting Eq. 10 into Eq. 9, we can rewrite the objective function of LINE(1st) as:

$$
\begin{aligned}
O_1 &= - \sum_{(i,j) \in \mathcal{E}} w_{ij} [\log \sigma(\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) + k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i})] \\
&= - \sum_{(i,j) \in \mathcal{E}} w_{ij} \log \sigma(\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) - \sum_{(i,j) \in \mathcal{E}} w_{ij} \cdot k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i}) \\
&= - \sum_{(i,j) \in \mathcal{E}} w_{ij} \log \sigma(\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) - \sum_{v_i \in \mathcal{V}} deg(v_i) \cdot k \cdot E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i})
\end{aligned}
\tag{11}
$$

To simplify the analysis, here we set the negative sampling distribution $P_n(v) \propto deg(v_n)$ [2]. Hence, the expectation term in Eq. 11 can be specified as follows:

$$
\begin{aligned}
E_{v_n \sim P_n(v)} \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i}) &= \sum_{v_n \in \mathcal{V}} \frac{deg(v_n)}{\sum_{v_n \in \mathcal{V}} deg(v_n)} \cdot \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i}) \\
&= \frac{deg(v_j)}{\sum_{v_n \in \mathcal{V}} deg(v_n)} \cdot \log \sigma(-\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) + \sum_{v_n \in \mathcal{V} \setminus \{v_j\}} \frac{deg(v_n)}{\sum_{v_n \in \mathcal{V}} deg(v_n)} \cdot \log \sigma(-\overrightarrow{v_n}^T \cdot \overrightarrow{v_i})
\end{aligned}
\tag{12}
$$

As each product $\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}$ is independent with others, we can gain the local objective for a specific (v_i, v_j) pair by combining Eqs. 11 and 12:

$$\ell(v_i, v_j) = w_{ij} \cdot \log \sigma(\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) + deg(v_i) \cdot k \cdot \frac{deg(v_j)}{\sum_{v_n \in \mathcal{V}} deg(v_n)} \cdot \log \sigma(-\overrightarrow{v_j}^T \cdot \overrightarrow{v_i}) \tag{13}$$

To minimize the objective function of LINE(1st) (i.e., Eq. 9), we must maximize $\ell(v_i, v_j)$. As such, we define $x = \overrightarrow{v_j}^T \cdot \overrightarrow{v_i}$ and get the derivative of $\ell(v_i, v_j)$ with respect to x:

$$\frac{\partial \ell}{\partial x} = w_{ij} \cdot \sigma(-x) - deg(v_i) \cdot k \cdot \frac{deg(v_j)}{\sum_{v_n \in \mathcal{V}} deg(v_n)} \cdot \sigma(x) \tag{14}$$

Comparing the derivative to zero, we have:

$$x = \overrightarrow{v_j}^T \cdot \overrightarrow{v_i} = \log \frac{w_{ij} \cdot \sum_{v_n \in \mathcal{V}} deg(v_n)}{deg(v_i) \cdot deg(v_j)} - \log k \tag{15}$$

Notably, in undirected networks, the PMI of (v_i, v_j) is $\log \frac{w_{ij} \cdot \sum_{v_n \in \mathcal{V}} deg(v_n)}{2 \cdot deg(v_i) \cdot deg(v_j)}$. Consequently, there exists the following relationship:

$$\log \frac{w_{ij} \cdot \sum_{v_n \in \mathcal{V}} deg(v_n)}{deg(v_i) \cdot deg(v_j)} = 2PMI(v_i, v_j) \tag{16}$$

---

[2]LINE(1st) sets the negative sampling distribution $P_n(v) \propto deg(v_n)^{3/4}$. In our proof, for simplicity, we replace it with $P_n(v) \propto deg(v_n)$ following [2].

4

Overall, therefore, we can characterize the matrix $M^{(1)}$ that LINE(1st) is actually factoring:

$$M_{ij}^{(1)} = \overrightarrow{v_j}^T \cdot \overrightarrow{v_i} = 2PMI(v_i, v_j) - \log k \qquad (17)$$

References

[1] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei.
Line: Large-scale information network embedding.
In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. ACM, 2015.

[2] Omer Levy and Yoav Goldberg.
Neural word embedding as implicit matrix factorization.
In *Advances in Neural Information Processing Systems 27*, pages 2177–2185. 2014.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.
Distributed representations of words and phrases and their compositionality.
In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[4] Kenneth Ward Church and Patrick Hanks.
Word association norms, mutual information, and lexicography.
Computational linguistics, 16(1):22–29, 1990.