Глубинное обучение в анализе графовых данных

3. Эмбеддинги

Высшая Школа Экономики 2022/2023

в предыдущих сериях…

Создание признаков

Графы - объекты сложные, нужно придумать как делать признаки

Как и для задач, разобрали признаки (feature engineering) для

- вершин
- связей
- графов

Признаки для вершин

- степень вершины
- центральность
- коэффициент кластеризации
- графлеты

Признаки для связей

- дистанция
- локальное пересечение
- глобальное пересечение

Признаки для графов

- графлетовые ядра
- WL ядра

Эмбеддинги

Мотивация

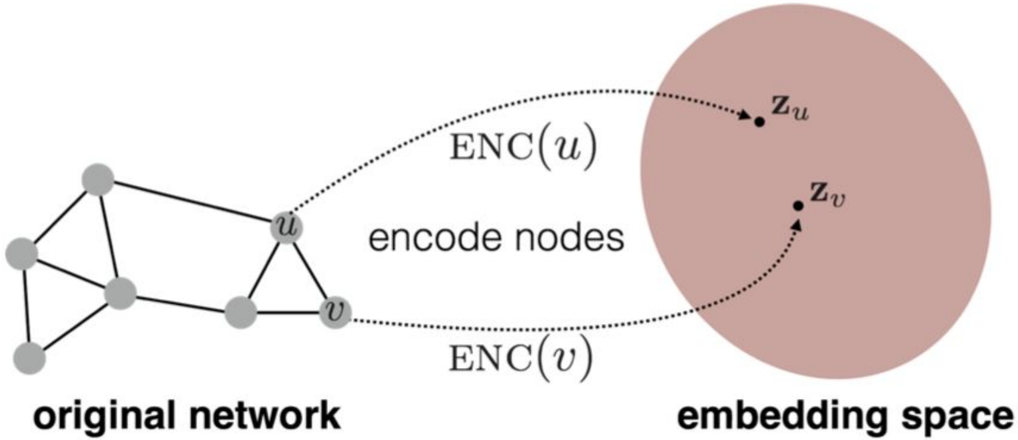
Вручную создавать признаки требует определенного погружения в задачу. На этапе построения не всегда понятно, какие точно стоит выбирать, какие помогут улучшить алгоритмы машинного обучения.

Нужно научиться выделять удобные признаки более “общими” способами. В этом нам смогут помочь **эмбеддинги**.

Эмбеддинги

Эмбединги вершин - векторы, суммаризующие позицию вершины в графе и локальное соседство

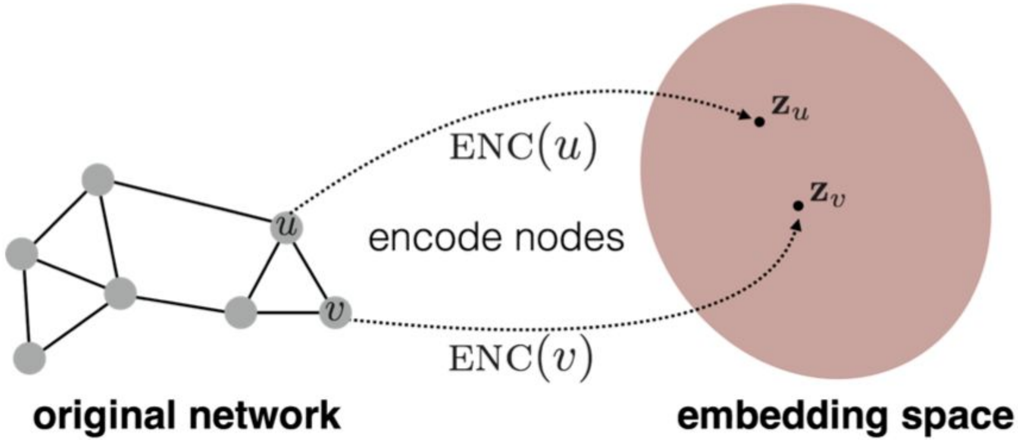
По сути хотим получить проекцию в пространство, где геометрические соотношения векторов будут отображать связи в оригинальном графе



Эмбеддинги

Перед нами стоит задача как-то закодировать вершины, чтобы похожесть в эмбеддинговом пространстве позволяла аппроксимировать похожесть в графе

similarity(u,v) ~ z_v ^T * z_u



Encoder

Encoder - функция, которая проецирует вершины графа в эмбеддинговое пространство

v -> z_v $\mathrm{ENC}: \mathcal{V} \to \mathbb{R}^d$

$$\mathrm{ENC}(v) = \mathbf{Z}[v]$$

Decoder

Decoder - функция, по представлению реконструирующая оригинал и связи

В простом случае парный декодировщик позволяет спрогнозировать связаны ли две вершины в графе

$$\mathrm{DEC} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$$

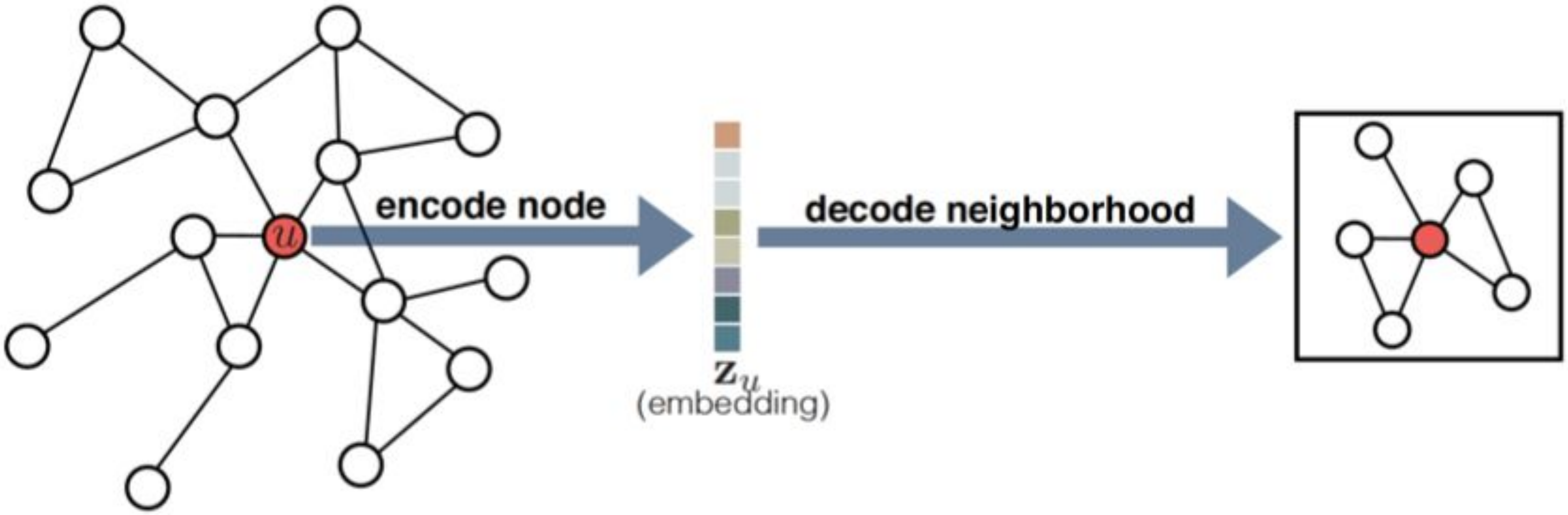
$$\mathrm{DEC}(\mathrm{ENC}(u), \mathrm{ENC}(v)) = \mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \approx \mathbf{S}[u, v].$$

Парадигма

1. encoder переводит вершины в пространство эмбеддингов
2. определяем каким-то образом похожесть вершин в графе
3. decoder переводит эмбеддинги в степень похожести
4. оптимизируем параметры encoder'а для достижения следующей цели

$$\text{similarity}(u, v) \approx \mathbf{z}_v^{\mathrm{T}} \mathbf{z}_u$$

Парадигма



Оптимизация

Задачу обучения декодировщировщика можно описать как задачу минимизации следующего функционала:

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \ell\left(\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v), \mathbf{S}[u, v]\right)$$

Подходы к созданию эмбеддингов

Основанные на факторизации матриц смежности

- Laplacian eigenmaps
- Inner-product methods

Laplacian eigenmaps

$$\text{DEC}(\mathbf{z}_u, \mathbf{z}_v) = \|\mathbf{z}_u - \mathbf{z}_v\|_2^2$$

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \text{DEC}(\mathbf{z}_u, \mathbf{z}_v) \cdot \mathbf{S}[u, v].$$

Inner-product methods

$$\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) = \mathbf{z}_u^\top \mathbf{z}_v$$

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \|\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) - \mathbf{S}[u, v]\|_2^2.$$

$$\mathcal{L} \approx \|\mathbf{Z}\mathbf{Z}^\top - \mathbf{S}\|_2^2$$

Обзор

Method	Decoder	Similarity measure	Loss function
Lap. Eigenmaps	$\lVert \mathbf{z}_u - \mathbf{z}_v \rVert_2^2$	general	$\text{DEC}(\mathbf{z}_u, \mathbf{z}_v) \cdot \mathbf{S}[u, v]$
Graph Fact.	$\mathbf{z}_u^\top \mathbf{z}_v$	$\mathbf{A}[u, v]$	$\lVert \text{DEC}(\mathbf{z}_u, \mathbf{z}_v) - \mathbf{S}[u, v] \rVert_2^2$
GraRep	$\mathbf{z}_u^\top \mathbf{z}_v$	$\mathbf{A}[u, v], ..., \mathbf{A}^k[u, v]$	$\lVert \text{DEC}(\mathbf{z}_u, \mathbf{z}_v) - \mathbf{S}[u, v] \rVert_2^2$
HOPE	$\mathbf{z}_u^\top \mathbf{z}_v$	general	$\lVert \text{DEC}(\mathbf{z}_u, \mathbf{z}_v) - \mathbf{S}[u, v] \rVert_2^2$
DeepWalk	$\frac{e^{\mathbf{z}_u^\top \mathbf{z}_v}}{\sum_{k \in \mathcal{V}} e^{\mathbf{z}_u^\top \mathbf{z}_k}}$	$p_{\mathcal{G}}(v \| u)$	$-\mathbf{S}[u, v] \log(\text{DEC}(\mathbf{z}_u, \mathbf{z}_v))$
node2vec	$\frac{e^{\mathbf{z}_u^\top \mathbf{z}_v}}{\sum_{k \in \mathcal{V}} e^{\mathbf{z}_u^\top \mathbf{z}_k}}$	$p_{\mathcal{G}}(v \| u)$ (biased)	$-\mathbf{S}[u, v] \log(\text{DEC}(\mathbf{z}_u, \mathbf{z}_v))$

Случайные блуждания

Обсужденные ранее методы как правило определяют S как некую полиномиальную функцию, и оптимизируется похожесть произведения двух эмбеддингов и S[u,v]

В последнее время успешными начинают быть методы, которые связаны с вероятностным подходом к похожести эмбеддингов, основанными на том, как часто вершины взаимно встречаются на случайных проходах

Случайные блуждания

$$\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \triangleq \frac{e^{\mathbf{z}_u^\top \mathbf{z}_v}}{\sum_{v_k \in \mathcal{V}} e^{\mathbf{z}_u^\top \mathbf{z}_k}}$$

Случайные блуждания

$$\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \triangleq \frac{e^{\mathbf{z}_u^\top \mathbf{z}_v}}{\sum_{v_k \in \mathcal{V}} e^{\mathbf{z}_u^\top \mathbf{z}_k}}$$

$$\approx p_{\mathcal{G},T}(v|u)$$

Случайные блуждания

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} -\log(\textsc{dec}(\mathbf{z}_u, \mathbf{z}_v))$$

node2vec (negative sampling)

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} -\log(\sigma(\mathbf{z}_u^\top \mathbf{z}_v)) - \gamma \mathbb{E}_{v_n \sim P_n(\mathcal{V})}[\log(-\sigma(\mathbf{z}_u^\top \mathbf{z}_{v_n}))].$$

LINE

$$\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) = \frac{1}{1 + e^{-\mathbf{z}_u^\top \mathbf{z}_v}}$$

Альтернативные подходы

- Можно совершать “прыжки” в блужданиях
- struct2vec

Проблемы

- каждая вершина в таком подходе рассматривается индивидуально
- не используют признаки вершин
- не способны работать с новыми вершинами

решим потом) (on the way to GNN)

Графы знаний (или multi-relational graphs)

Граф - $G = (V, E)$, V - вершины, E - ребра

В обычном графе $e = (u, v)$

В графе знаний $e = (u, t, v)$

В общем случае на графах знаний решается задача предсказания пропущенных связей, но бывают и задачи классификации вершин

RESCAL

$$\mathrm{DEC}(u, \tau, v) = \mathbf{z}_u^\top \mathbf{R}_\tau \mathbf{z}_v$$

$$\begin{aligned}\mathcal{L} &= \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{\tau \in \mathcal{R}} \|\mathrm{DEC}(u, \tau, v) - \mathcal{A}[u, \tau, v]\|^2 \\ &= \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{\tau \in \mathcal{R}} \|\mathbf{z}_u^\top \mathbf{R}_\tau \mathbf{z}_v - \mathcal{A}[u, \tau, v]\|^2,\end{aligned}$$

Стандартные функции потерь

- cross-entropy with negative sampling
- max-margin loss

Cross-entropy with negative sampling

$$\mathcal{L} = \sum_{(u,\tau,v)\in\mathcal{E}} -\log(\sigma(\text{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v))) - \gamma \mathbb{E}_{v_n \sim P_{n,u}(\mathcal{V})} \left[\log\left(\sigma\left(-\text{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_{v_n})\right)\right)\right]$$

$$\mathcal{L} = \sum_{(u,\tau,v)\in\mathcal{E}} \left(-\log(\sigma(\text{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v))) - \sum_{v_n \in \mathcal{P}_{n,u}} \left[\log\left(\sigma\left(-\text{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_{v_n})\right)\right)\right] \right)$$

Max-margin loss

$$\mathcal{L} = \sum_{(u,\tau,v)\in\mathcal{E}} \sum_{v_n\in\mathcal{P}_{n,u}} \max(0, -\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) + \mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_{v_n}) + \Delta)$$

Обзор

Name	Decoder	Relation Parameters
RESCAL	$\mathbf{z}_u^\top \mathbf{R}_\tau \mathbf{z}_v$	$\mathbf{R}_\tau \in \mathbb{R}^{d \times d}$
TransE	$-\|\mathbf{z}_u + \mathbf{r}_\tau - \mathbf{z}_v\|$	$\mathbf{r}_\tau \in \mathbb{R}^d$
TransX	$-\|g_{1,\tau}(\mathbf{z}_u) + \mathbf{r}_\tau - g_{2,\tau}(\mathbf{z}_v)\|$	$\mathbf{r}_\tau \in \mathbb{R}^d, g_{1,\tau}, g_{2,\tau} \in \mathbb{R}^d \to \mathbb{R}^d$
DistMult	$< \mathbf{z}_u, \mathbf{r}_\tau, \mathbf{z}_v >$	$\mathbf{r}_\tau \in \mathbb{R}^d$
ComplEx	$\mathrm{Re}(< \mathbf{z}_u, \mathbf{r}_\tau, \bar{\mathbf{z}}_v >)$	$\mathbf{r}_\tau \in \mathbb{C}^d$
RotatE	$-\|\mathbf{z}_u \circ \mathbf{r}_\tau - \mathbf{z}_v\|$	$\mathbf{r}_\tau \in \mathbb{C}^d$

Translational Decoders

$$\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) = -\|\mathbf{z}_u + \mathbf{r}_\tau - \mathbf{z}_v\|$$

$$\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) = -\|g_{1,\tau}(\mathbf{z}_u) + \mathbf{r}_\tau - g_{2,\tau}(\mathbf{z}_v)\|$$

$$\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) = -\|(\mathbf{z}_u - \mathbf{w}_r^\top \mathbf{z}_u \mathbf{w}_r) + \mathbf{r}_\tau - (\mathbf{z}_v - \mathbf{w}_r^\top \mathbf{z}_v \mathbf{w}_r)\|$$

Multi-linear

$$\begin{aligned}
\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) &=< \mathbf{z}_u, \mathbf{r}_\tau, \mathbf{z}_v > \\
&= \sum_{i=1}^{d} \mathbf{z}_u[i] \times \mathbf{r}_\tau[i] \times \mathbf{z}_v[i] \\
&=< \mathbf{z}_v, \mathbf{r}_\tau, \mathbf{z}_u > \\
&= \mathrm{DEC}(\mathbf{z}_v, \tau, \mathbf{z}_u).
\end{aligned}$$

Complex

$$\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) = \mathrm{Re}(< \mathbf{z}_u, \mathbf{r}_\tau, \bar{\mathbf{z}}_v >)$$

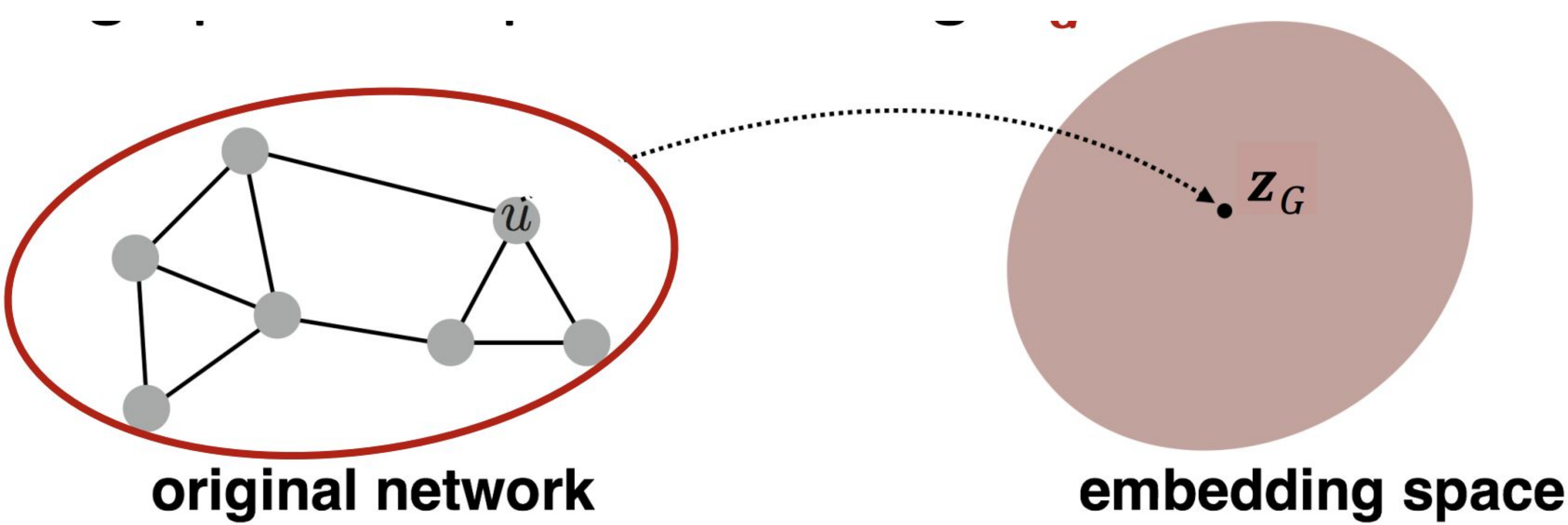
$$= \mathrm{Re}(\sum_{i=1}^{d} \mathbf{z}_u[i] \times \mathbf{r}_\tau[i] \times \bar{\mathbf{z}}_v[j])$$

$$\mathrm{DEC}(\mathbf{z}_u, \tau, \mathbf{z}_v) = -\|\mathbf{z}_u \circ \mathbf{r}_\tau - \mathbf{z}_v\|,$$

Свойства

- симметрия (асимметрия)
- инверсионность
- композитность

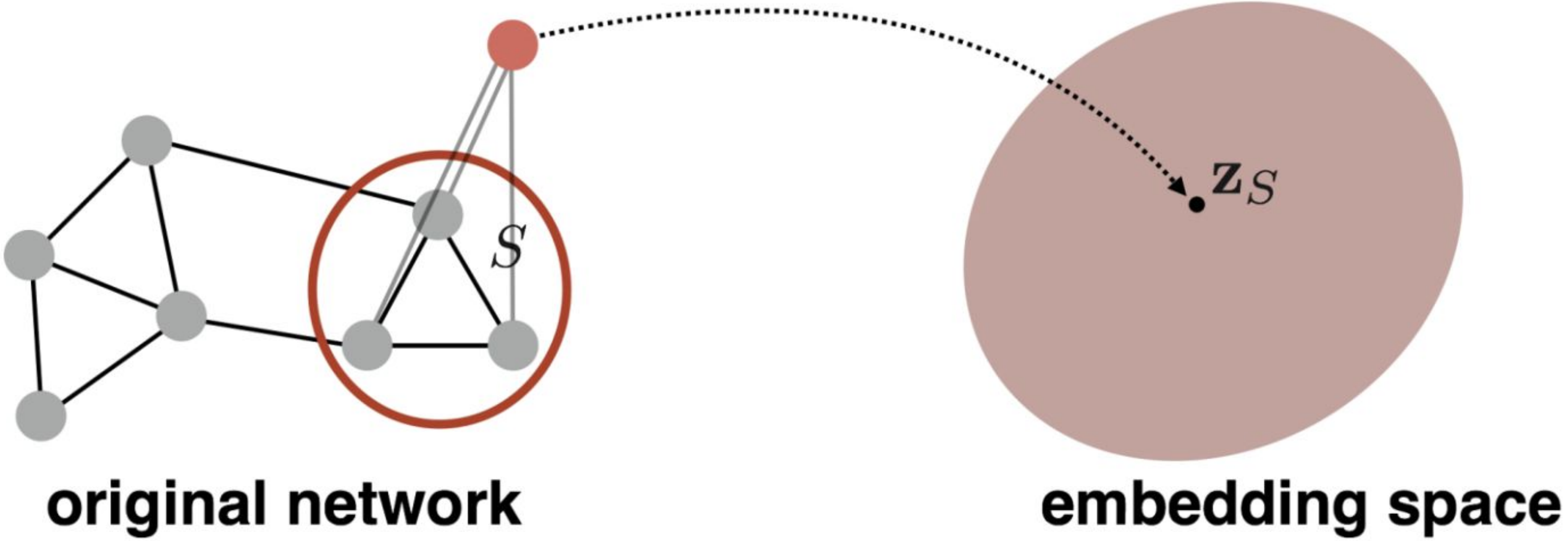
Эмбеддинги графов



Среднее

$$\mathbf{z}_G = \sum_{v \in G} \mathbf{z}_v$$

Виртуальная вершина



Анонимные случайные блуждания

1. Сэмплируем анонимные случайные блуждания и отображаем граф как частоты появления случайных блужданий
2. Обучаем эмбеддинги графа вместе с анонимными случайными блужданиями