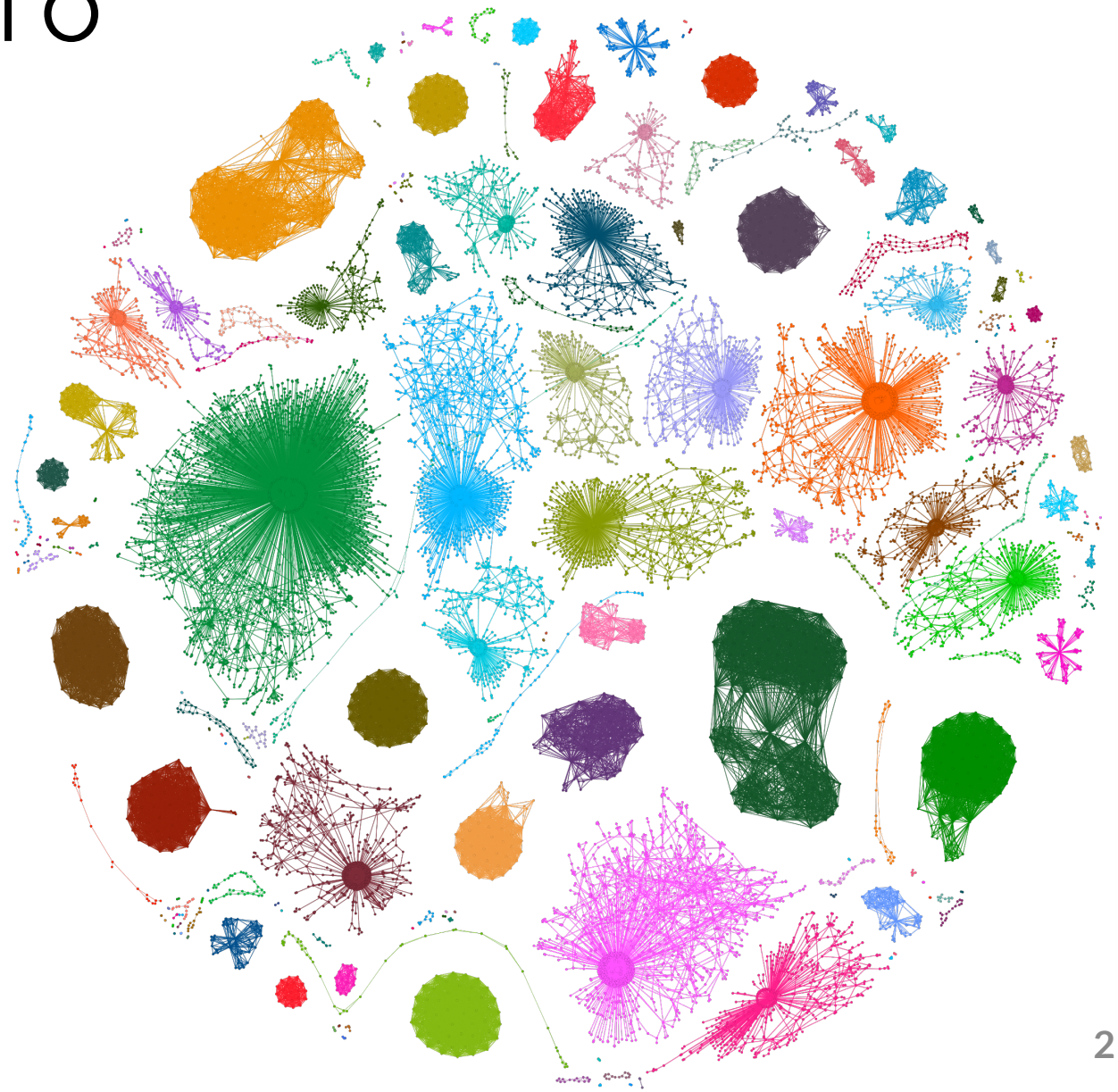
Эмбеддинги графов без учителя

Антон Цицулин

Графов в мире много

Разные **домены**:

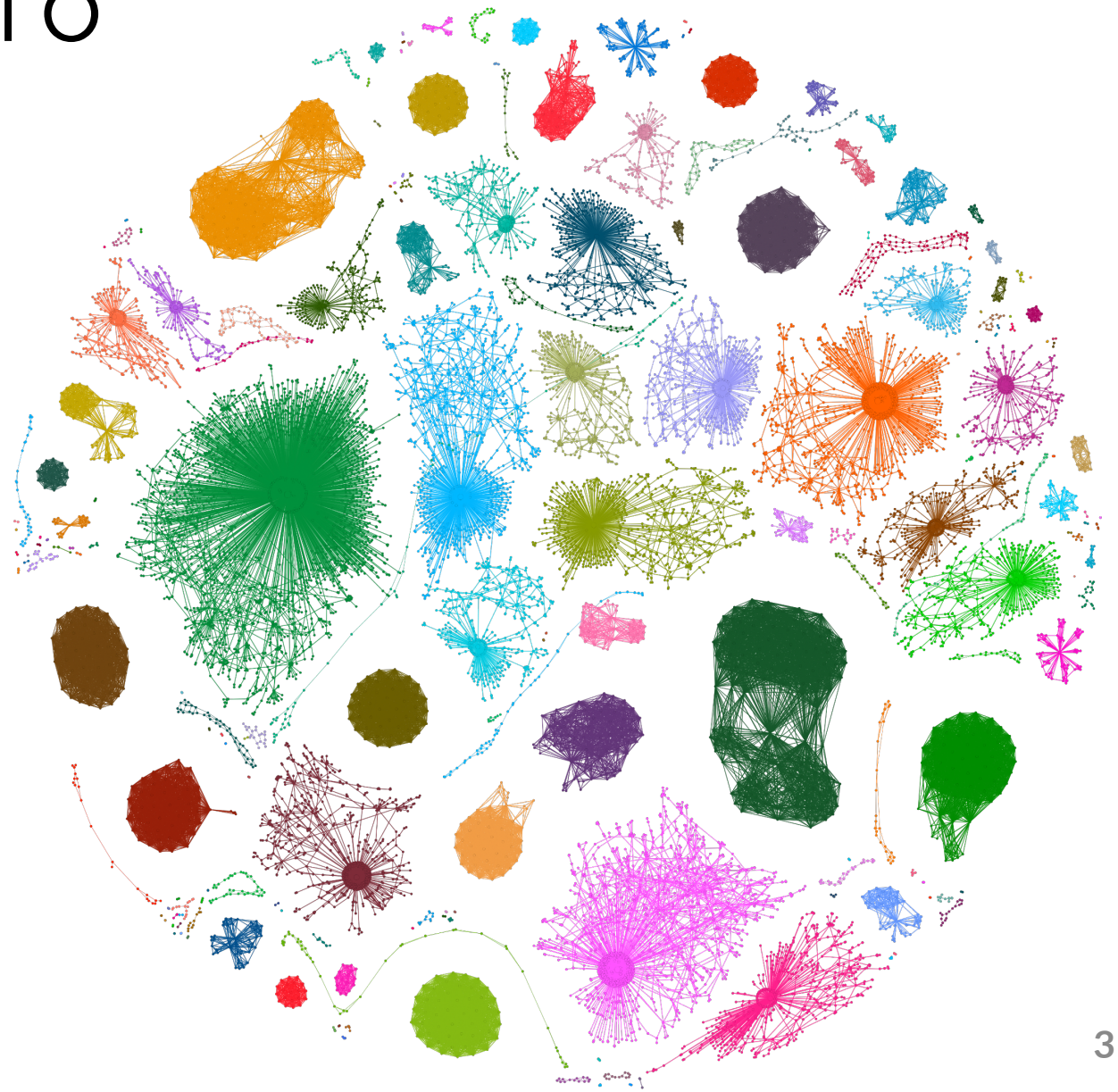
- Соцсети
- Биологические
- Транспортные сети
- …



Графов в мире много

Разные **типы графов**:

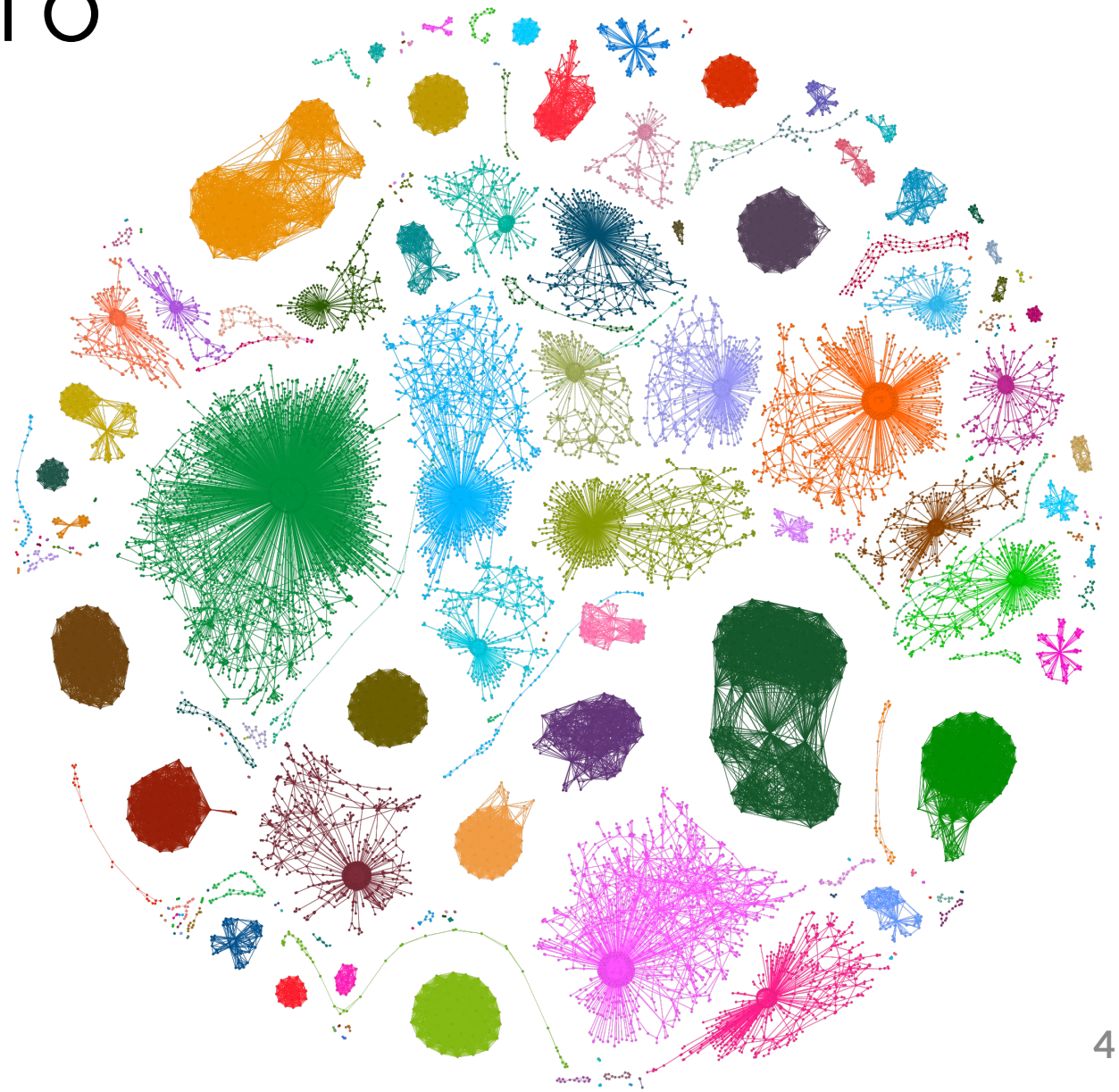
- (Не)направленные
- (Не)взвешенные
- С атрибутами
- Гетерогенные
- …



Графов в мире много

Разные **уровни анализа**:

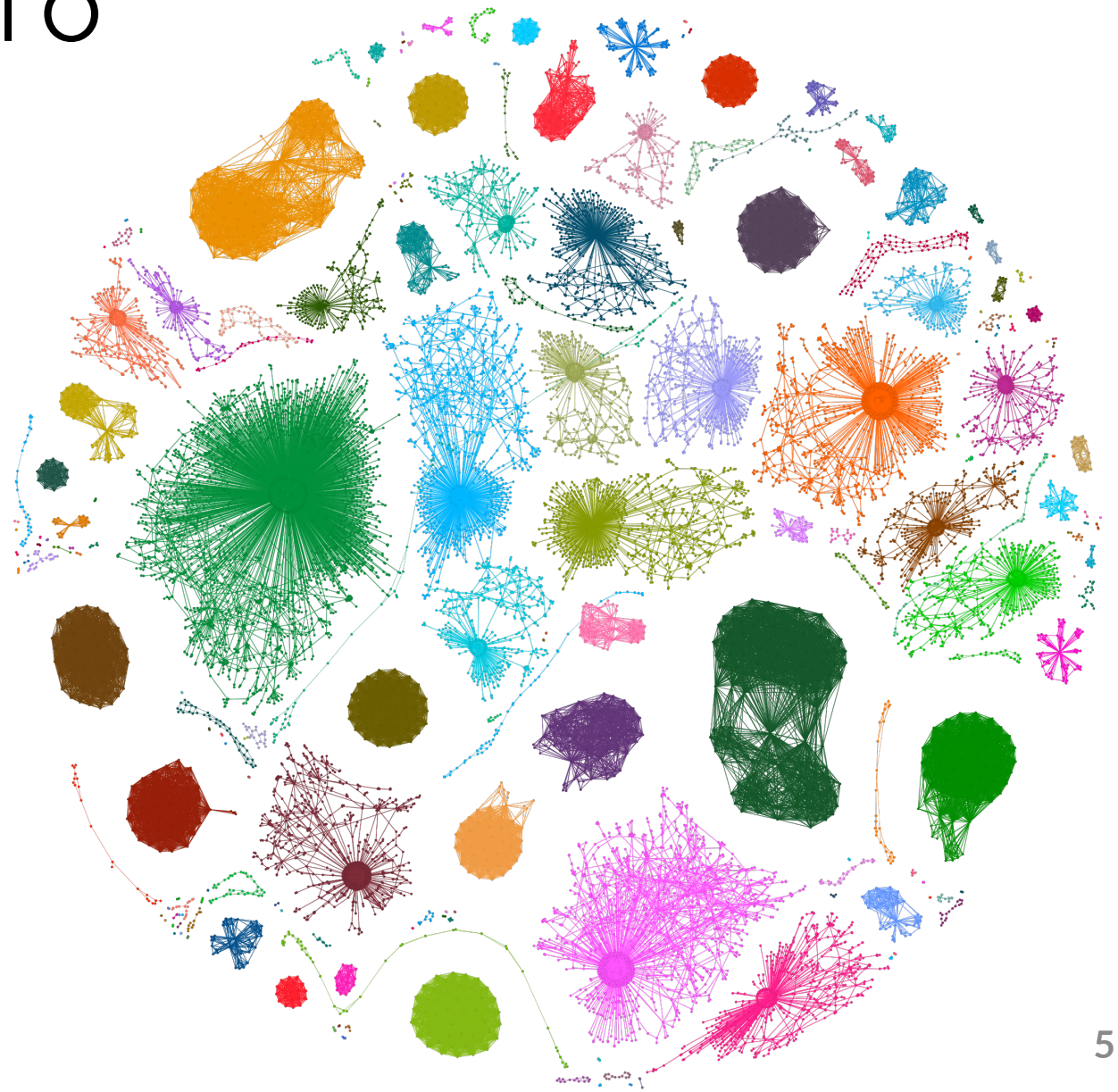
- Вершины
- Рёбра
- Подграфы
- Графы целиком
- …



Графов в мире много

Разные **задачи**:

- Классификация
- Кластеризация
- Поиск аномалий
- …



Графов в мире много

Домены	Типы	Уровни	Задачи
Соцсети Биологические Транспортные сети	(Не)направленные (Не)взвешенные С атрибутами Гетерогенные	Вершины Рёбра Подграфы Графы целиком	Классификация Кластеризация Поиск аномалий

Графов в мире много

Домены

Соцсети
Биологические
Транспортные сети

Типы

(Не)направленные
(Не)взвешенные
С атрибутами
Гетерогенные

Уровни

Вершины
Рёбра
Подграфы
Графы целиком

Задачи

Классификация
Кластеризация
Поиск аномалий

↓

Эмбеддинги

Графов в мире много

Домены

Соцсети
Биологические
Транспортные сети

Типы

(Не)направленные
(Не)взвешенные
С атрибутами
Гетерогенные

Уровни

Вершины
Рёбра
Подграфы
Графы целиком

Задачи

Классификация
Кластеризация
Поиск аномалий

↓

Эмбеддинги

Графов в мире много

Домены

Соцсети
Биологические
Транспортные сети

Типы

(Не)направленные
(Не)взвешенные
С атрибутами
Гетерогенные

Уровни

Вершины
Рёбра
Подграфы
Графы целиком

Задачи

Классификация
Кластеризация
Поиск аномалий

↓

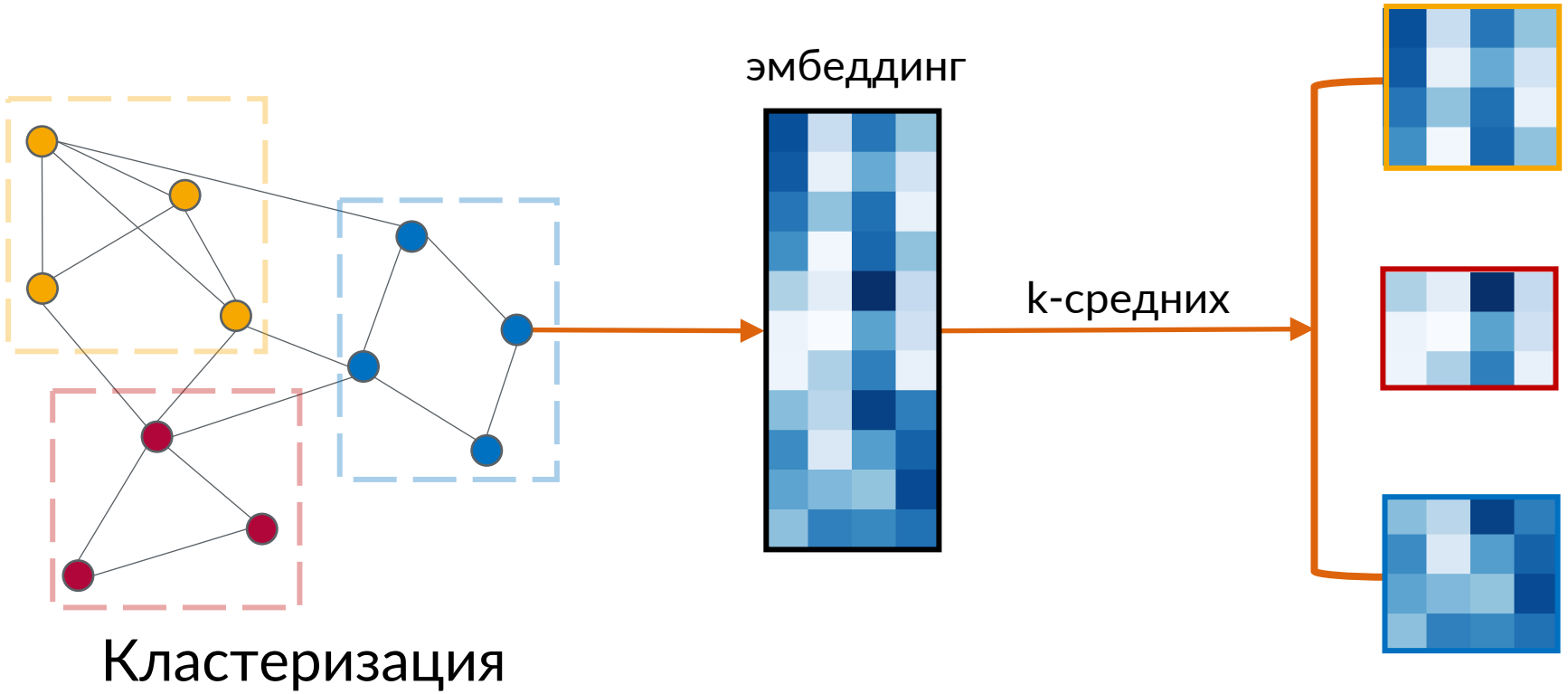
Эмбеддинги

Графов в мире много

Домены	Типы	Уровни	Задачи
Соцсети Биологические Транспортные сети	(Не)направленные (Не)взвешенные С атрибутами Гетерогенные	Вершины Рёбра Подграфы Графы целиком	Классификация Кластеризация Поиск аномалий ↓ Эмбеддинги

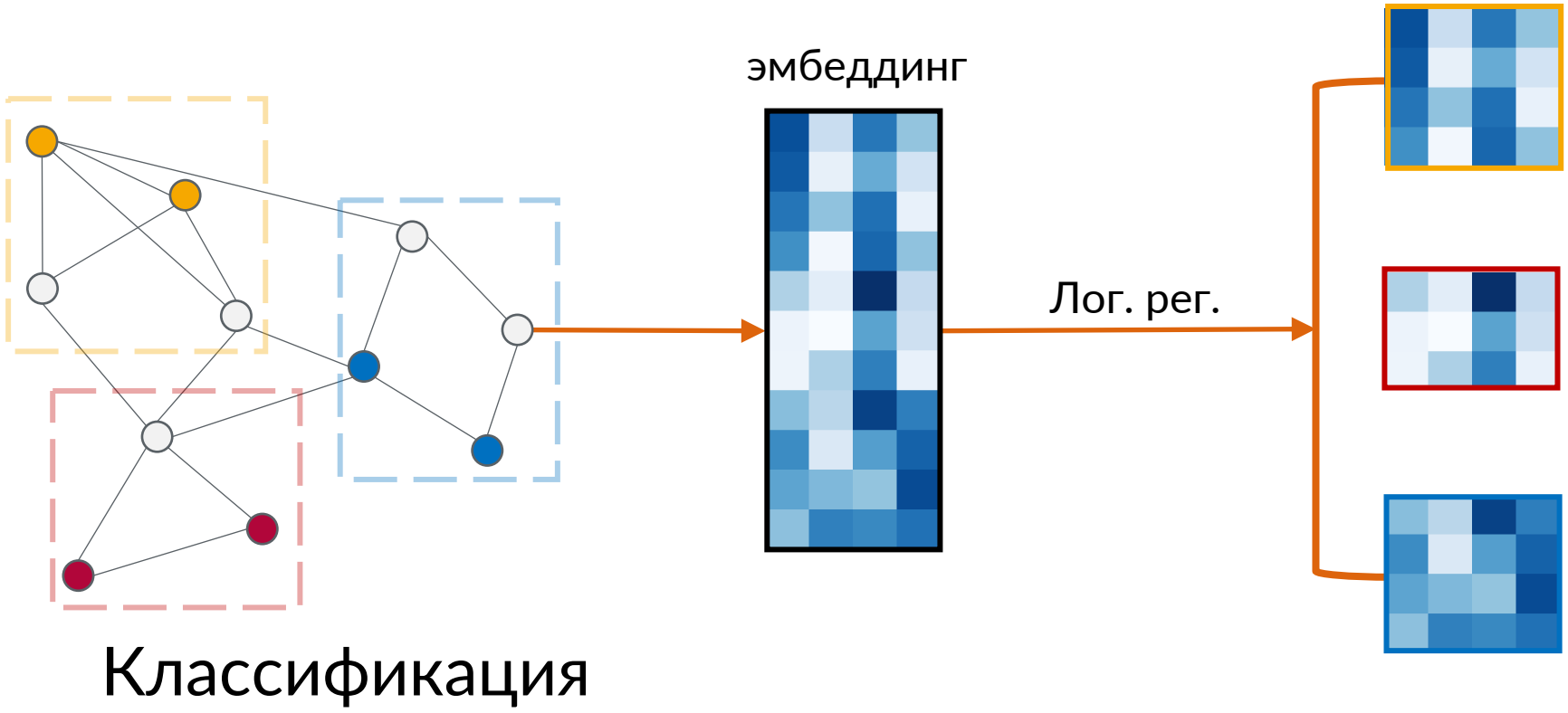
Почему эмбеддинги?

У нас есть хорошие алгоритмы для анализа векторных данных ...



Почему эмбеддинги?

У нас есть хорошие алгоритмы для анализа векторных данных ...

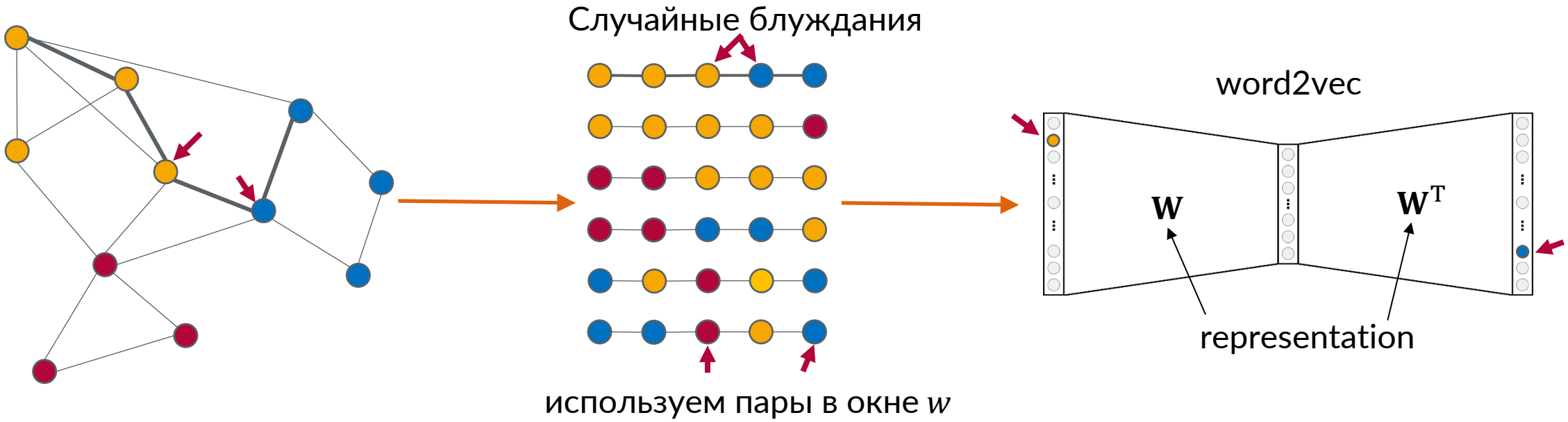


Часть 1: графы без атрибутов

DeepWalk

“Вершины в случ. блужданиях≈словам в предложениях → бахнем word2vec”

Начинаем γ блужданий длины t из каждой вершины



DeepWalk: асимптотика и практика

На практике, $\gamma = 80$, $t = 80$, $w = 10$, получаем $80 * 80 * n$ “текста”

NB: никогда не меняем w

Если снижаете w, увеличивайте γ и t

Как правильно тюнить параметры мы не знаем :(

DeepWalk: асимптотика и практика

Код на питоне делает все блуждания и вызывает word2vec

Я написал версию на C++ которая так не делает

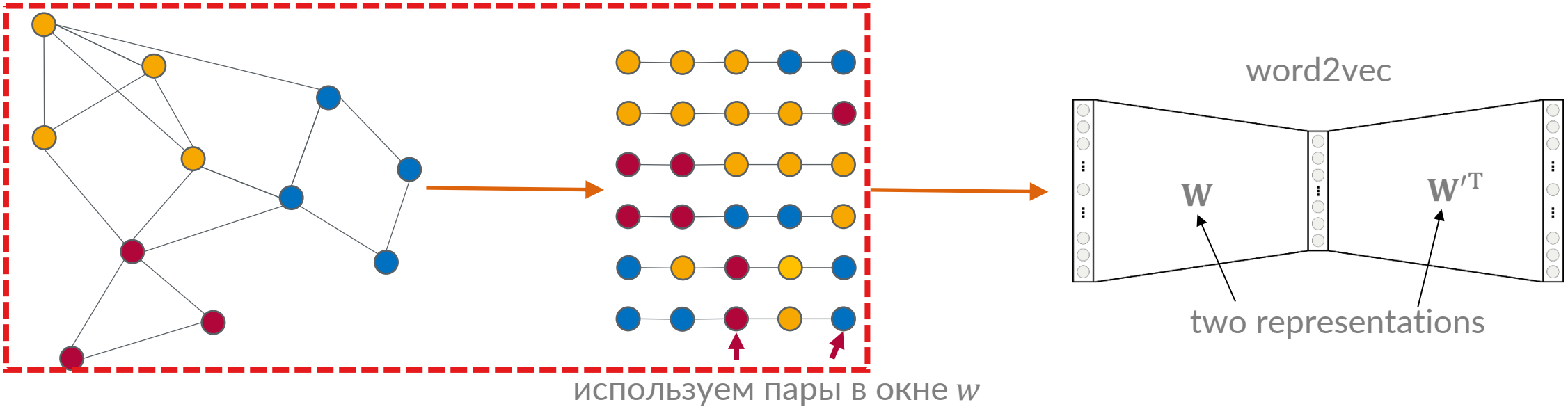
Эпоха делается за $O(d * \log n)$:(

На практике: ~3M вершин

Node2vec

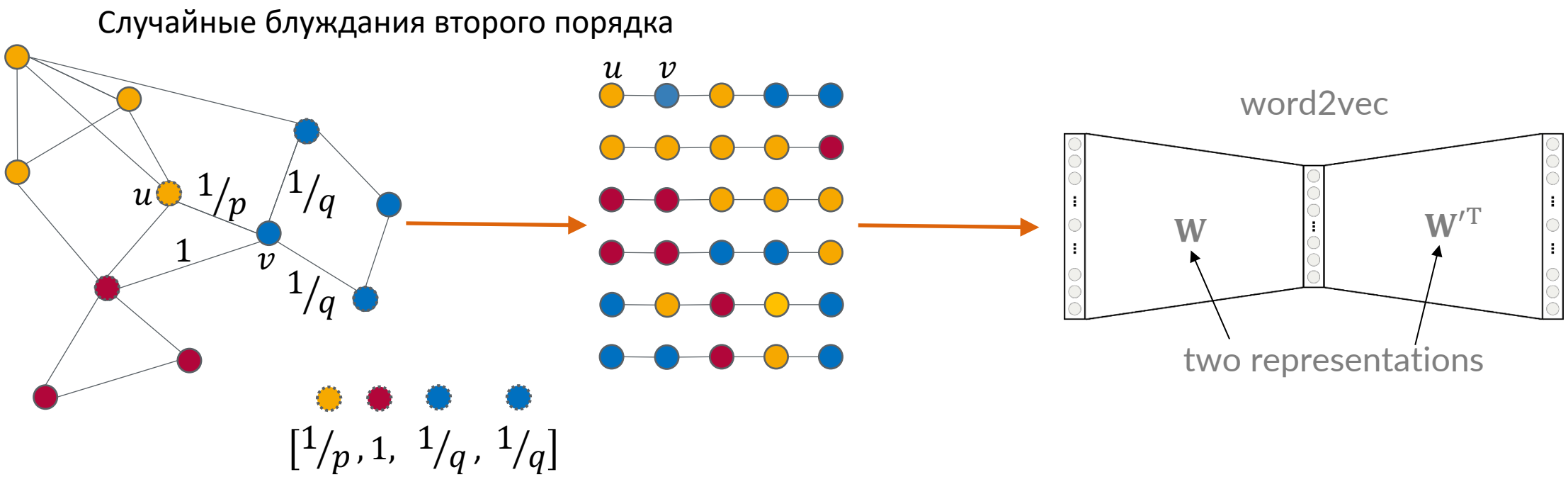
“Давайте-ка добавим еще два параметра (p, q) в DeepWalk”

Начинаем γ блужданий длины t из каждой вершины



Node2vec

“Давайте-ка добавим еще два параметра (p, q) в DeepWalk”



Node2vec: scalable feature learning for networks. Grover & Leskovec, KDD 2016

Node2vec: миф №1

Миф: параметры (p, q) отвечают за BFS и DFS

Реальность: параметры (p, q) отвечают за трегульники $\approx$ кластеры

Низкий $q \rightarrow$ блуждатель ходит между кластерами

Высокий $q \rightarrow$ блуждатель ходит внутри кластеров

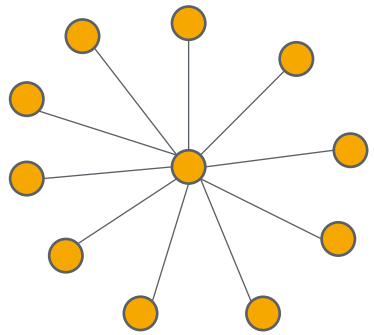
Node2vec: миф №2

Миф: node2vec быстро работает

Реальность: случайные блуждания второго порядка - $O(n^2)$

Худший случай для графов-звёзд

(или любых графов с вершинами с высокой степенью)



Node2vec: scalable feature learning for networks. Grover & Leskovec, KDD 2016

Node2vec: асимптотика и практика

NB: сравнения в статье обманчивы ($\gamma = 10$ для всех)

В статье, $\gamma = 10$, $t = 80$, $w = 10$, перебор (p, q)

$\gamma = 10$ даёт плохие результаты, используйте $\gamma = 80$

Подбор (p, q) не даёт прироста на большинстве графов

Node2vec: асимптотика и практика

Код на питоне делает все блуждания и вызывает word2vec

Я написал версию на C++ которая так не делает

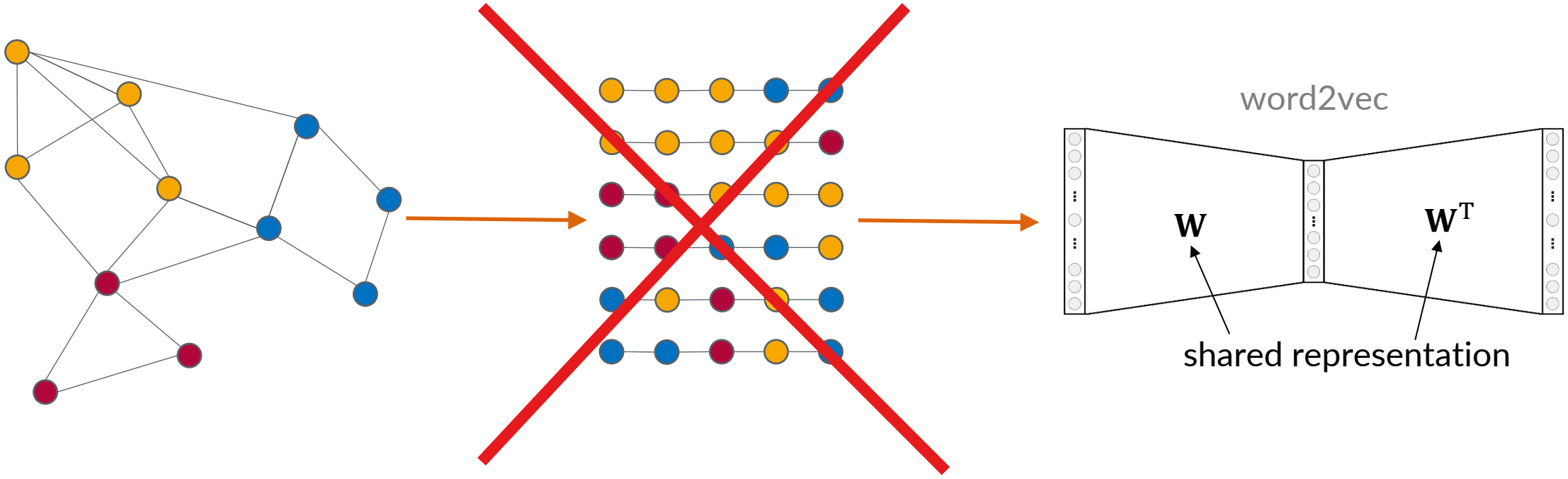
Препроцессинг до $O(n^2)$:(

Эпоха делается за $O(d)$:)

На практике: ~500т вершин если повезло с графом, если нет, ~50т

VERSE

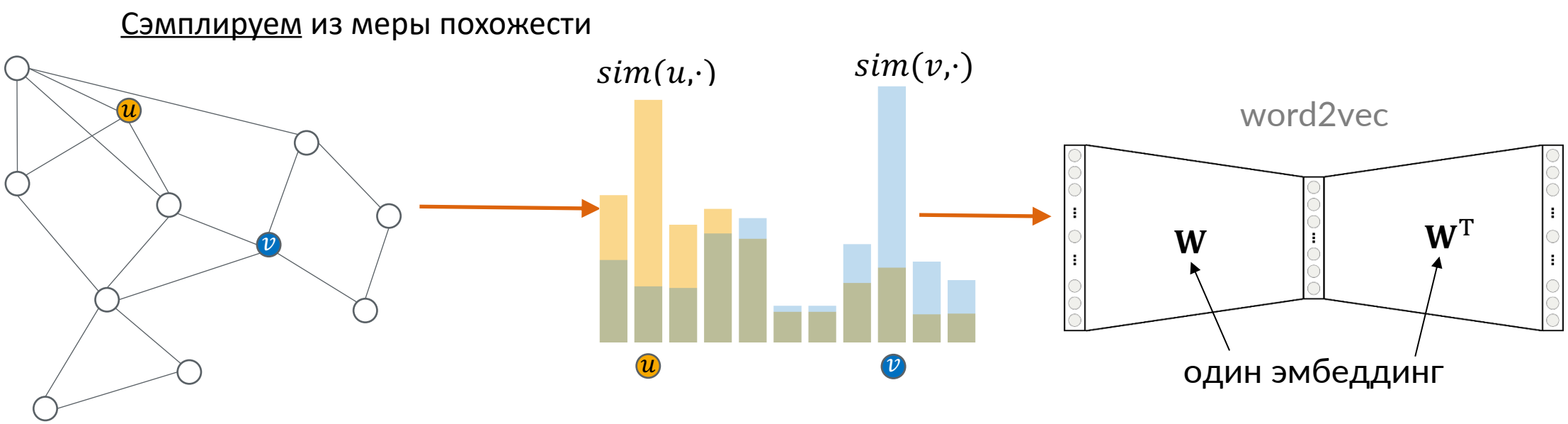
“Случайные блуждания – мера похожести вершин”



VERSE: Versatile graph embeddings from similarity measures. Tsitsulin et al., WWW 2018

VERSE

“Случайные блуждания – мера похожести вершин”

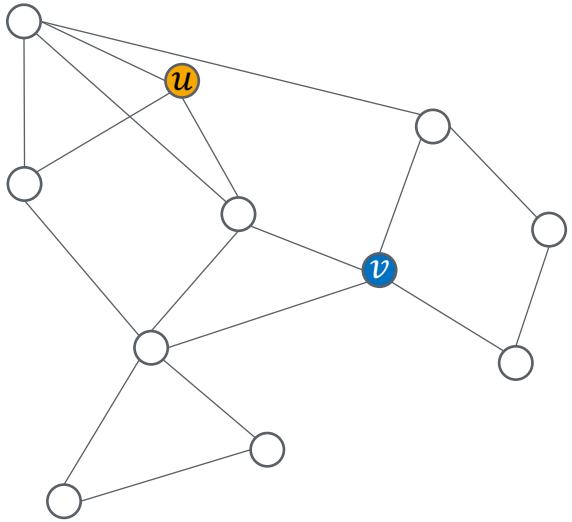


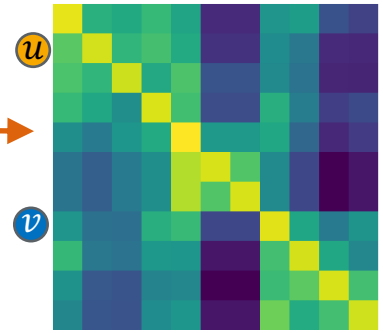
VERSE: Versatile graph embeddings from similarity measures. Tsitsulin et al., WWW 2018

VERSE

“Random walks define a similarity distribution”

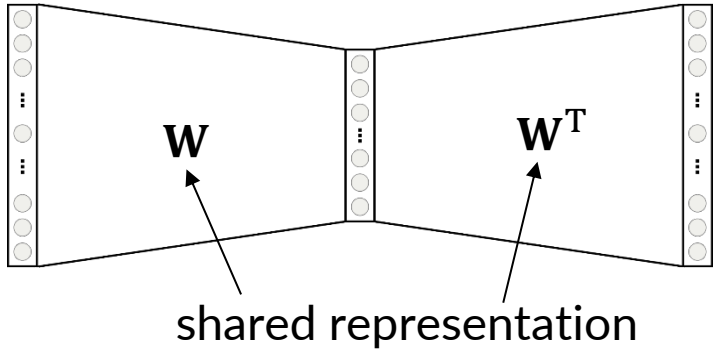
Сэмплируем из меры похожести





Матрица попарной похожести $\mathbf{M}$
не материализуется

Predictions with negative sampling



VERSE: интерпретация DeepWalk

Блуждания в DeepWalk ~= Personalized PageRank

Параметр PPR $\alpha = \frac{w-2}{w+1}$ для w в DeepWalk

Мы можем мерить качество эмбеддингов :)

1 параметр вместо 3 или 5

VERSE: асимптотика и практика

Простой и быстрый алгоритм, хорош на предсказании рёбер :)

Эпоха делается за $O(d)$:)

Код на C++ работает хорошо

На практике: ~10M вершин

Эмбеддинги для рёбер

Operator	Result
Average	$(\mathbf{a} + \mathbf{b})/2$
Concat	$[\mathbf{a}_1, \dots, \mathbf{a}_d, \mathbf{b}_1, \dots, \mathbf{b}_d]$
Hadamard	$[\mathbf{a}_1 * \mathbf{b}_1, \dots, \mathbf{a}_d * \mathbf{b}_d]$
Weighted L1	$[\lvert\mathbf{a}_1 - \mathbf{b}_1\rvert, \dots, \lvert\mathbf{a}_d - \mathbf{b}_d\rvert]$
Weighted L2	$[(\mathbf{a}_1 - \mathbf{b}_1)^2, \dots, (\mathbf{a}_d - \mathbf{b}_d)^2]$

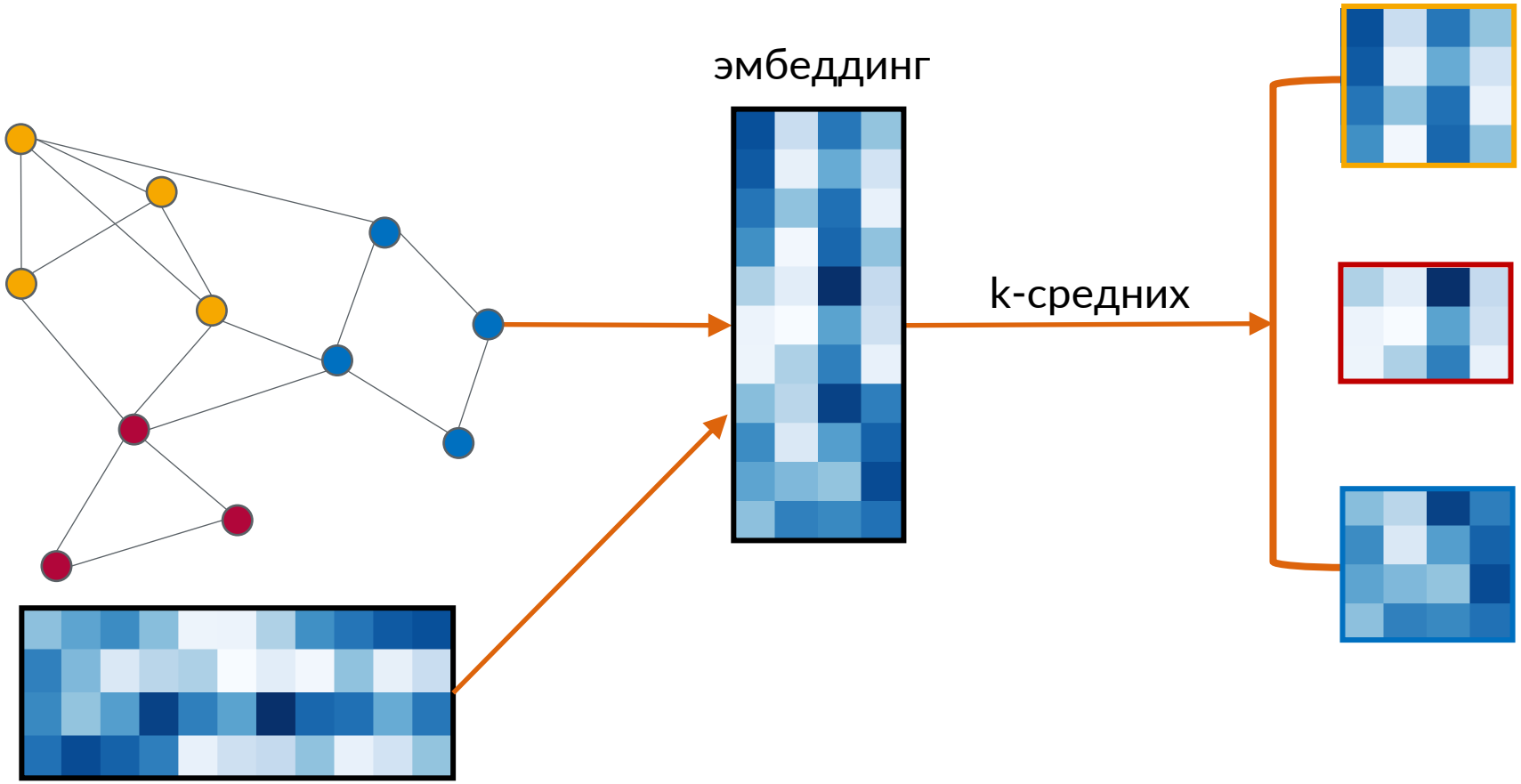
Table 2.3: Vector operators used for link-prediction task for each $u, v \in V$ and corresponding embeddings $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$.

NB: оператор выбираем в зависимости от алгоритма

Часть 1: графы с атрибутами

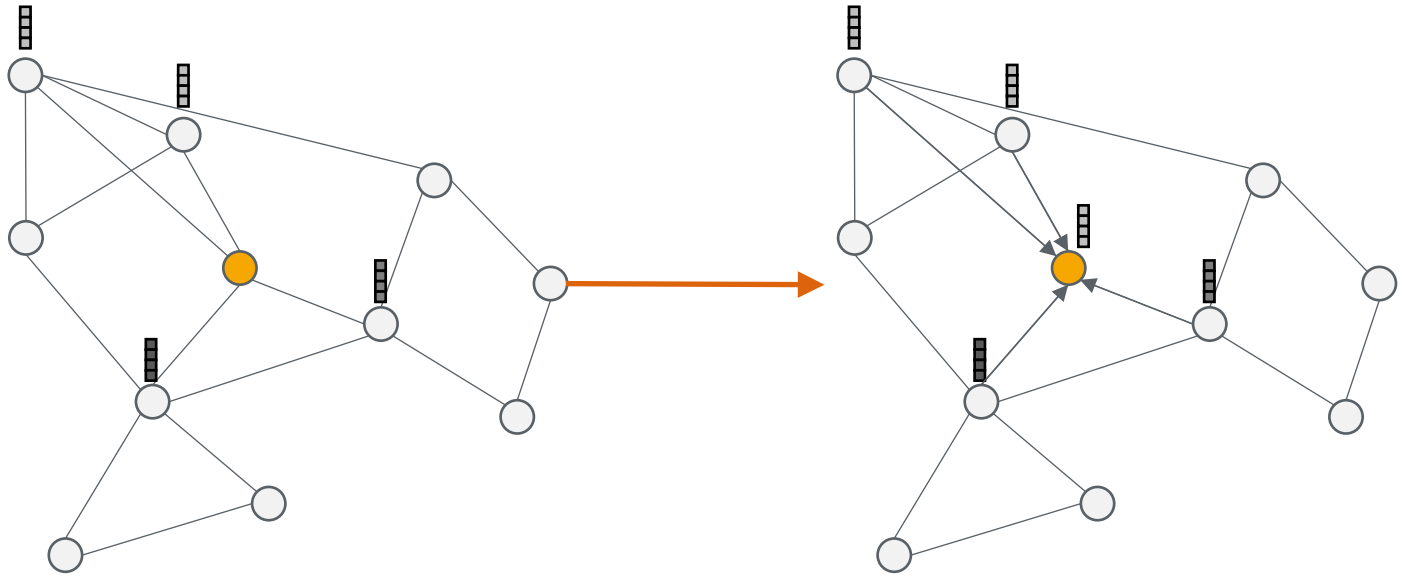
Графы с атрибутами

На каждой вершине есть некоторый вектор с фичами



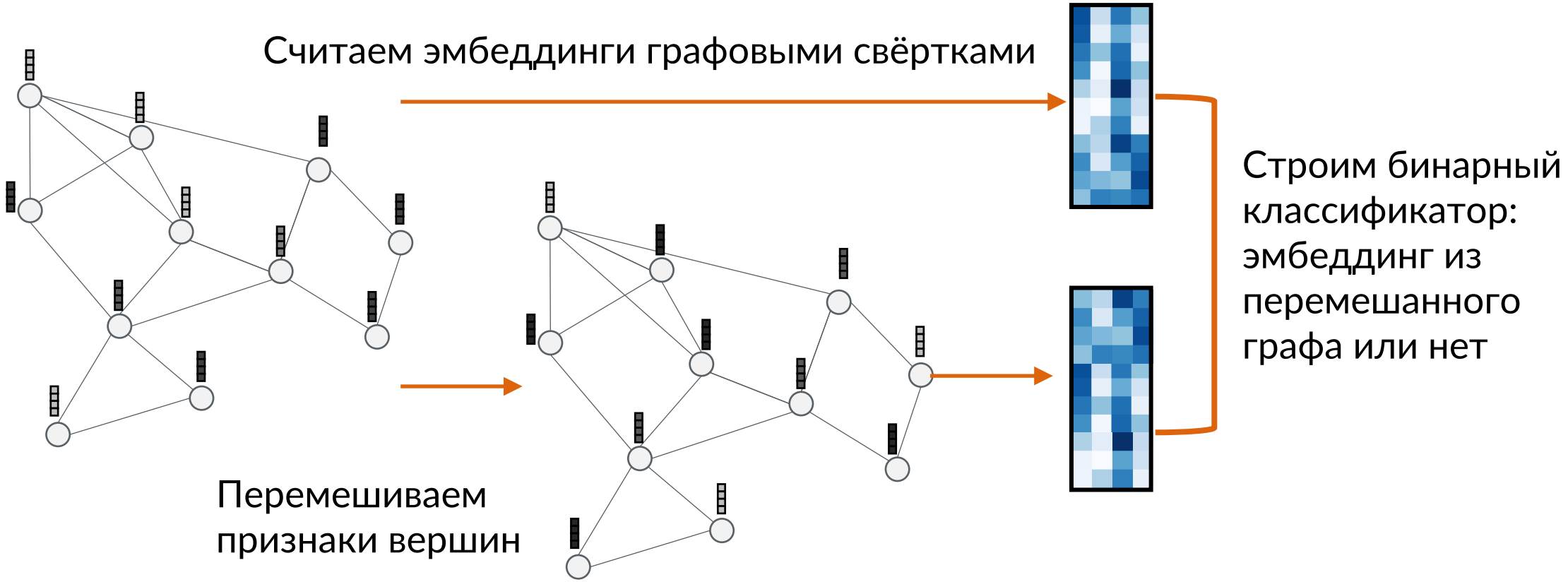
Графовые свёртки

Скрытое представление вершины = среднее представление соседей



Deep Graph Infomax

“Учимся отличать граф от него же с перемешанными признаками”



Deep Graph Infomax. Veličković et al., ICLR 2019.

DGI: асимптотика и практика

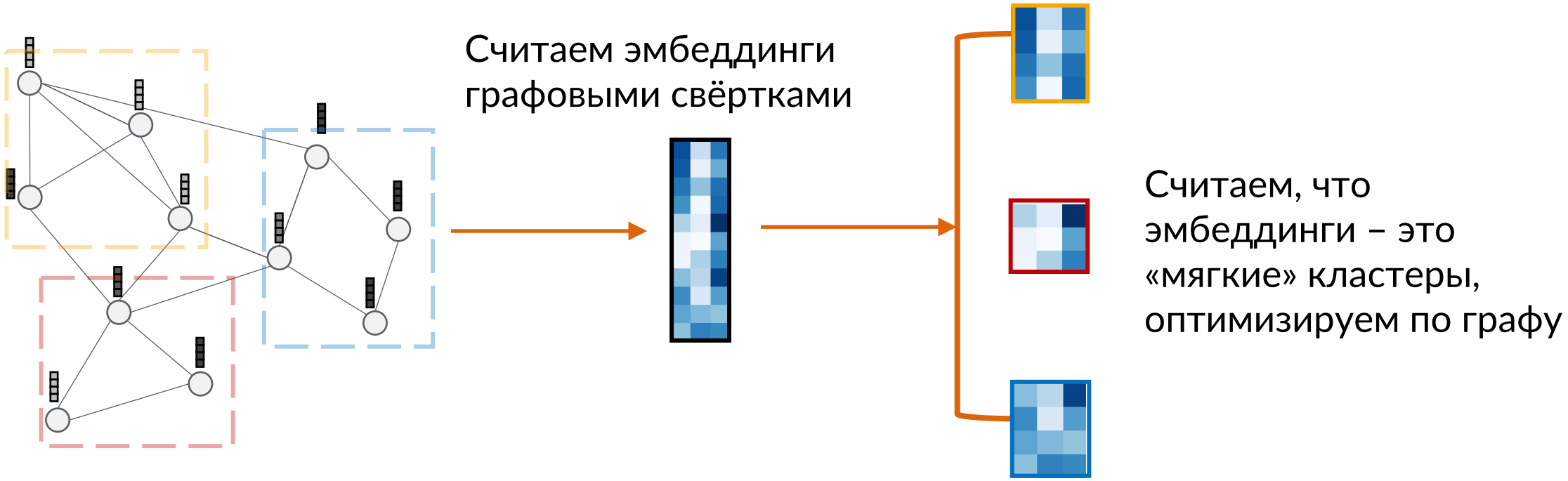
Можно делать для больших графов :)

«Размазывает» информацию по графу

Нетривиальное обучение – нужен early stopping :(

Deep Modularity Networks

“Учимся кластеризовывать граф”



Считаем, что эмбеддинги – это «мягкие» кластеры, оптимизируем по графу

Graph Clustering with Graph Neural Networks. Tsitsulin et al., arXiv 2020.

DMoN: асимптотика и практика

Можно делать для средних графов (1M) :/

Группирует вершины графа по графу

Простое обучение

Остались вопросы?

Twitter	twitter.com/tsitsulin_	
Сайт	tsitsul.in/talks/datastart	← презентация
Пишите	anton@tsitsul.in	