Методы получения репрезентативной выборки пользователей социальных сетей

Гомзин Андрей Геннадьевич

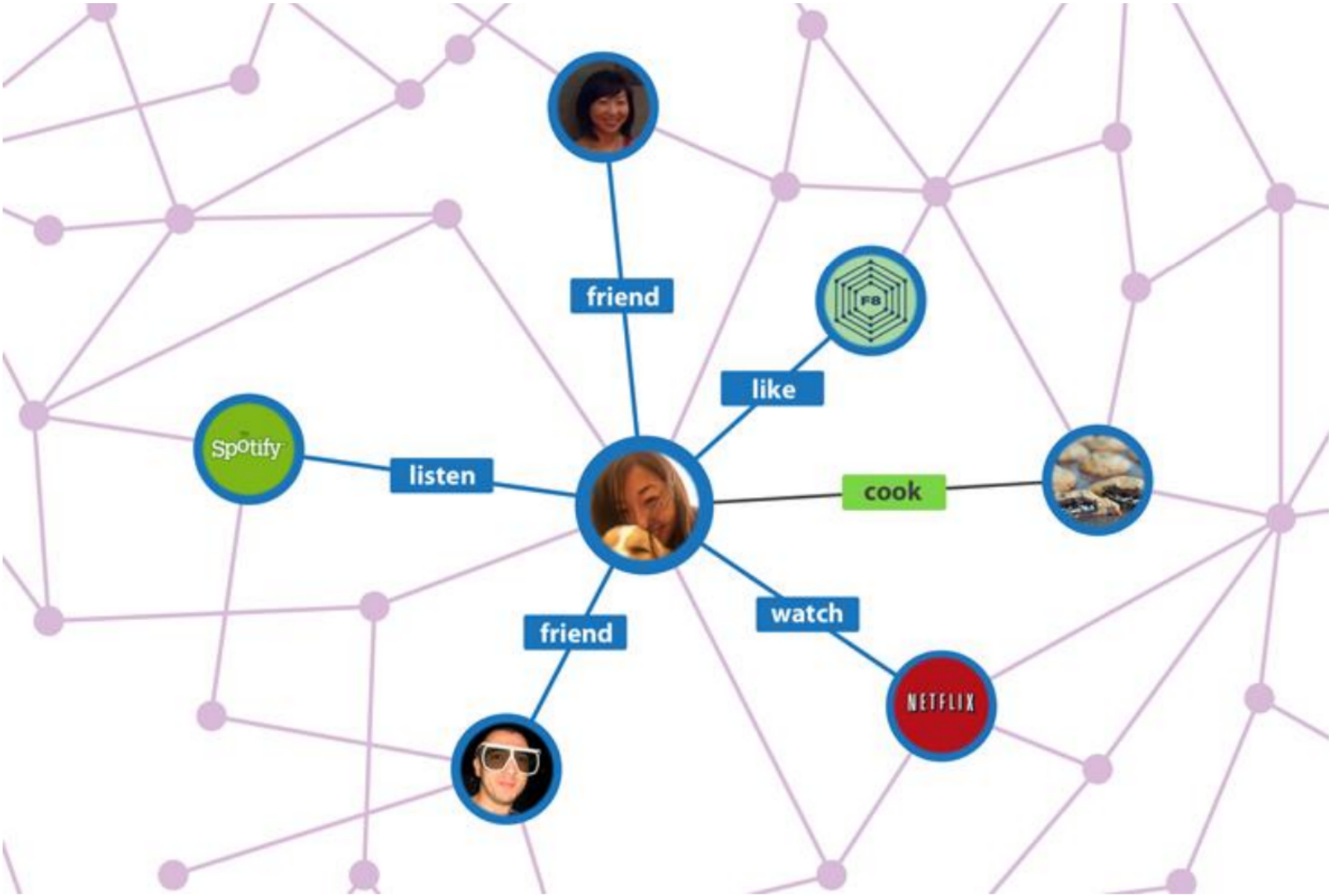
ИСП РАН

18 марта 2014 г.

План

1. Введение
2. Критерии репрезентативности выборки
3. Методы сэмплинга
4. MODIS Crawler

Социальная сеть



Объекты (в т.ч. пользователи), связи между ними

Социальные сети. Количество пользователей

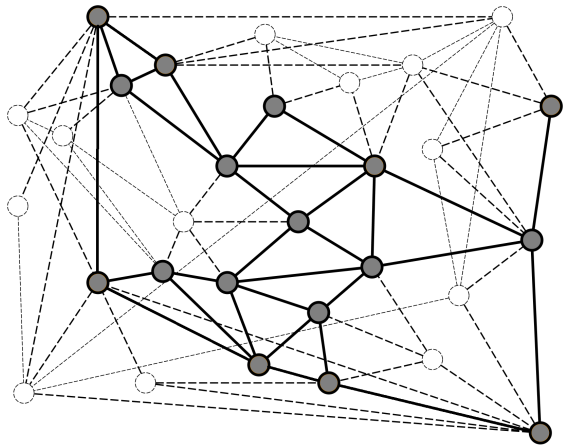
- *Facebook*
  - **1 миллиард** зарегистрированных пользователей
- *Twitter*
  - **500 млн** всего, **140 млн** активных пользователей
- *ВКонтакте*
  - **230 млн** зарегистрированных пользователей, **60 млн** посетителей в сутки
- *Одноклассники*
  - **205 млн** пользователей, более **40 млн** посетителей в сутки
- *LastFM*
  - **40 млн** зарегистрированных пользователей

$\Rightarrow$ скачать всю социальную сеть практически невозможно

$\Rightarrow$ будем скачивать небольшую ее часть

Сэмплинг

- **Сэмплинг** – процесс получения (репрезентативной) выборки объектов и связей социальной сети
- Предполагается, что информация о вершине становится известной только после ее включения в выборку (сэмпл)
- Упрощенная модель социальной сети:
  - Гомогенная сеть:
    - ★ Один тип вершин (пользователи)
    - ★ Один тип ребер (отношение дружбы/фолловинг)
  - Для каждого пользователя скачивается вся информация о нем:
    - ★ Профиль
    - ★ Список друзей/фолловеров
    - ★ Связанные объекты другого типа (сообщения, мультимедиа)

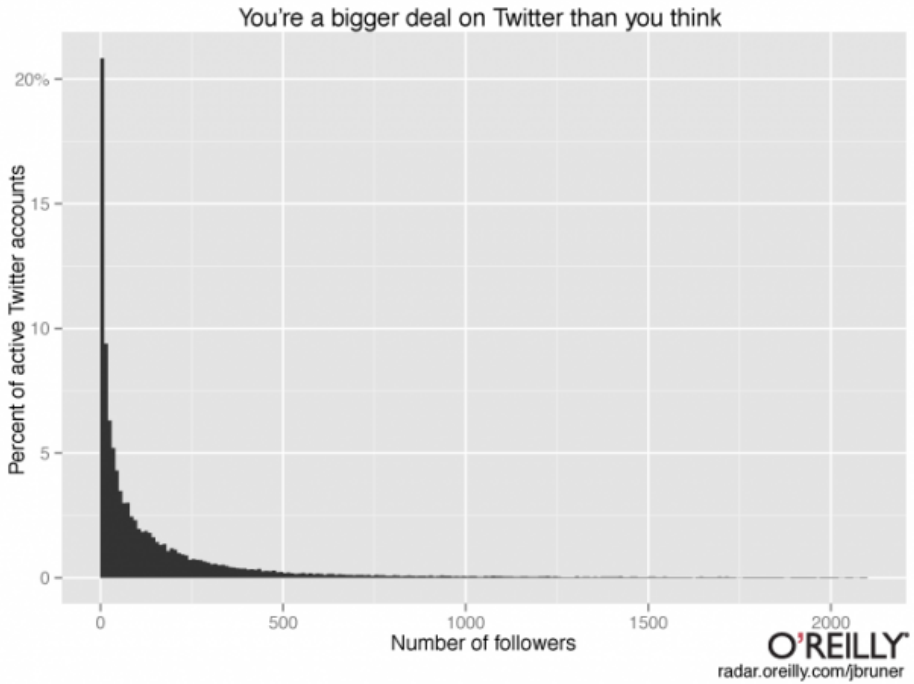


Репрезентативность выборки (1-2)

Полученная выборка должна иметь те же свойства (распределения), что и оригинальный граф:

1. In-degree distribution: количество пользователей с заданным числом входящих ребер
2. Out-degree distribution: количество пользователей с заданным числом исходящих ребер

Степенной закон



Репрезентативность выборки (3-4)

Полученная выборка должна иметь те же свойства (распределения), что и оригинальный граф:

3. Распределение размеров компонент слабой связности (WCC):
   - Множество вершин составляет компоненту *слабой* связности, если между любой парой вершин существует *ненаправленный* путь.
4. Распределение размеров компонент сильной связности (SCC):
   - Множество вершин составляет компоненту *сильной* связности, если между любой парой вершин u и v существуют *направленные* пути из u в v и из v в u.

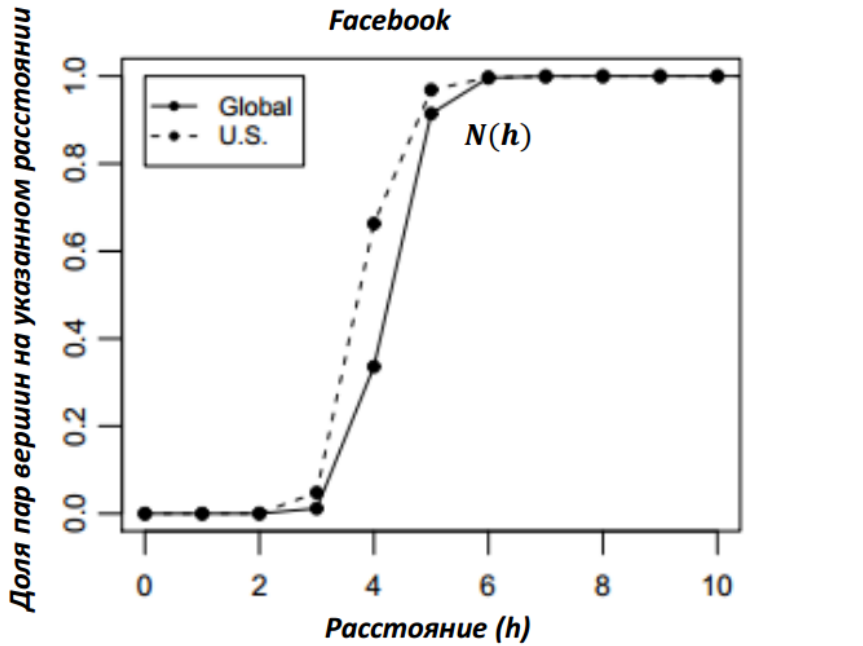
Репрезентативность выборки (5-6)

Полученная выборка должна иметь те же свойства (распределения), что и оригинальный граф:

5. Hop-plot: количество $N(h)$ пар вершин, достижимых за h шагов
6. Hop-plot на компоненте слабой связности (WCC) максимального размера

Закон "малого мира"

N(h) показывает процент пар вершин, достижимых за **h** шагов.



Репрезентативность выборки (7-8)

Полученная выборка должна иметь те же свойства (распределения), что и оригинальный граф:

7. Распределение первого левого сингулярного вектора *матрицы смежности графа*, отсортированного по убыванию
8. Распределение сингулярных чисел *матрицы смежности графа*, отсортированные по убыванию

Неотрицательное вещественное число σ называется **сингулярным числом** матрицы $M \in R^{m \times n}$ если и только если существуют два вектора единичной длины $u \in R^m$ и $v \in R^n$ такие, что:
$Mv = \sigma u$ и $M^* u = \sigma v$.
u и v – соответственно, **левый** и **правый сингулярные вектора**.

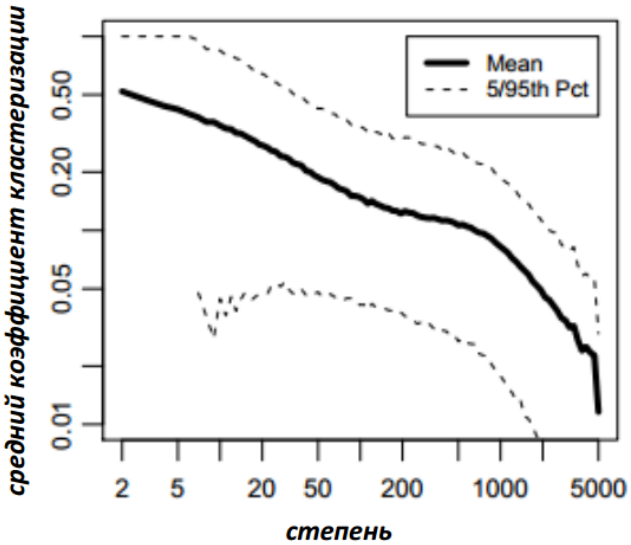
Репрезентативность выборки (9)

Полученная выборка должна иметь те же свойства (распределения), что и оригинальный граф:

9. Распределение коэффициента кластеризации

- Пусть k – количество друзей пользователя
- Максимальное число ребер между k друзьями: $\frac{k(k-1)}{2}$
- **Коэффициент кластеризации** – доля существующих ребер из максимально возможного

Пример для Facebook:



*Для пользователя **Facebook** со степенью **100** средний коэффициент кластеризации **0.14**, т.е. **14%** его друзей состоят в друзьях друг у друга.*

Методы сэмплинга

Классификация:

- Выбор случайной вершины
- Выбор случайного ребра
- Исследование социальной сети

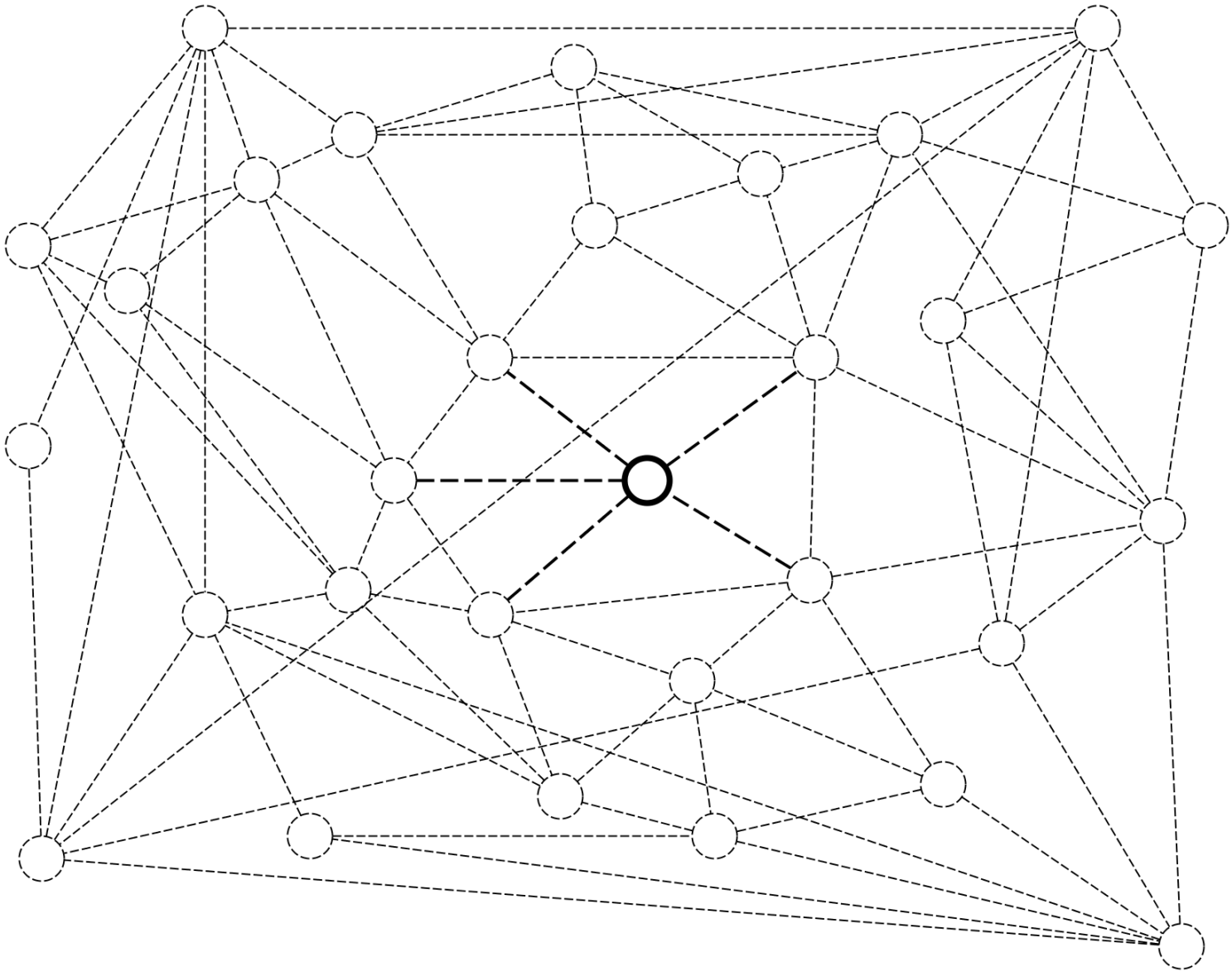
Методы сэмплинга. Выбор случайной вершины

Предполагается, что идентификаторы вершин графа известны

- Random Node (RN): Выбор из равномерного распределения
  - Сэмпл не удовлетворяет степенному закону
- Random Degree Node (RDN): Выбор из распределения степени вершины
  - Предполагается, что степени вершин известны
- Random PageRank Node (RPN): Выбор из распределения PageRank
  - Предполагается, что PageRank вершин известны

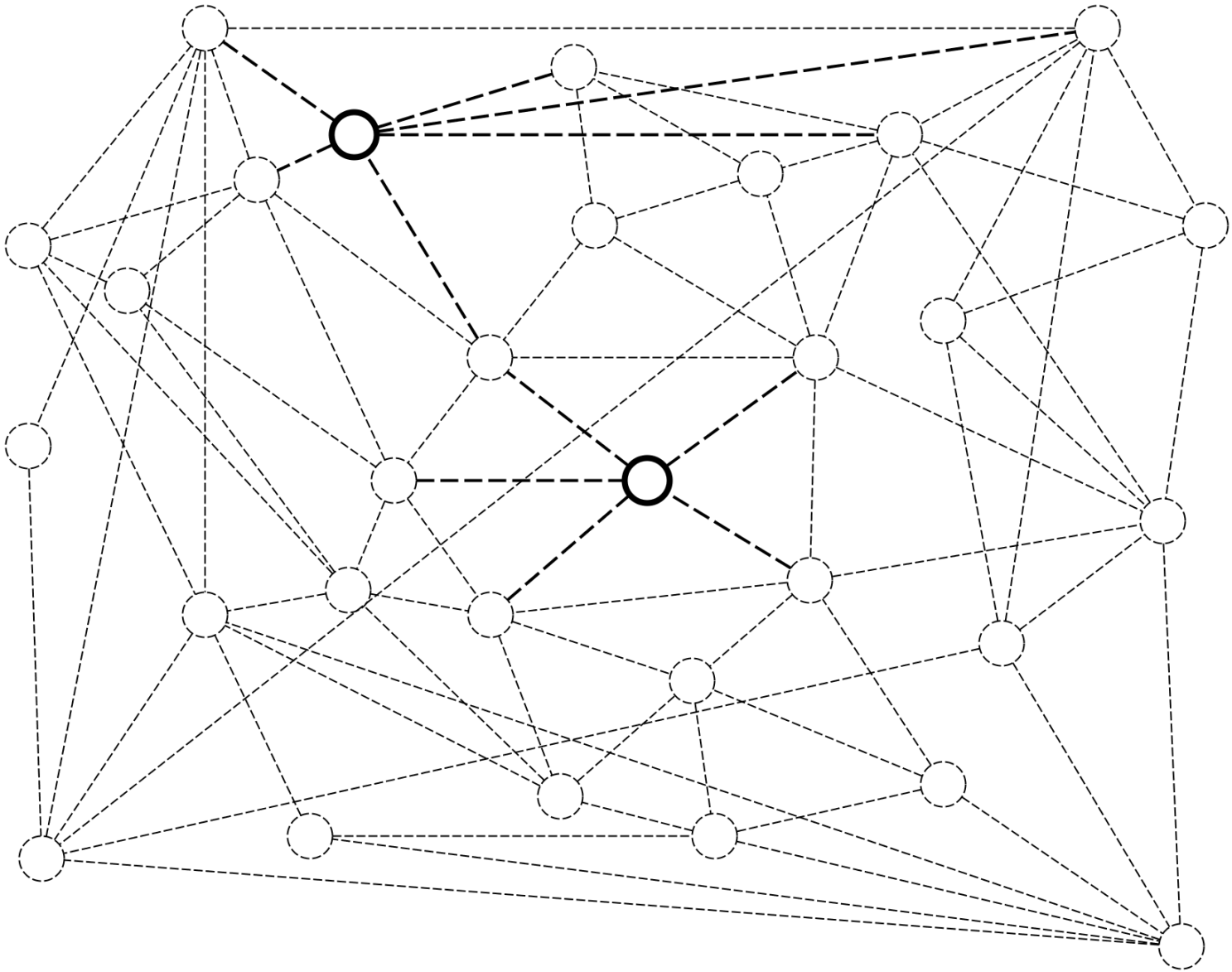
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 1



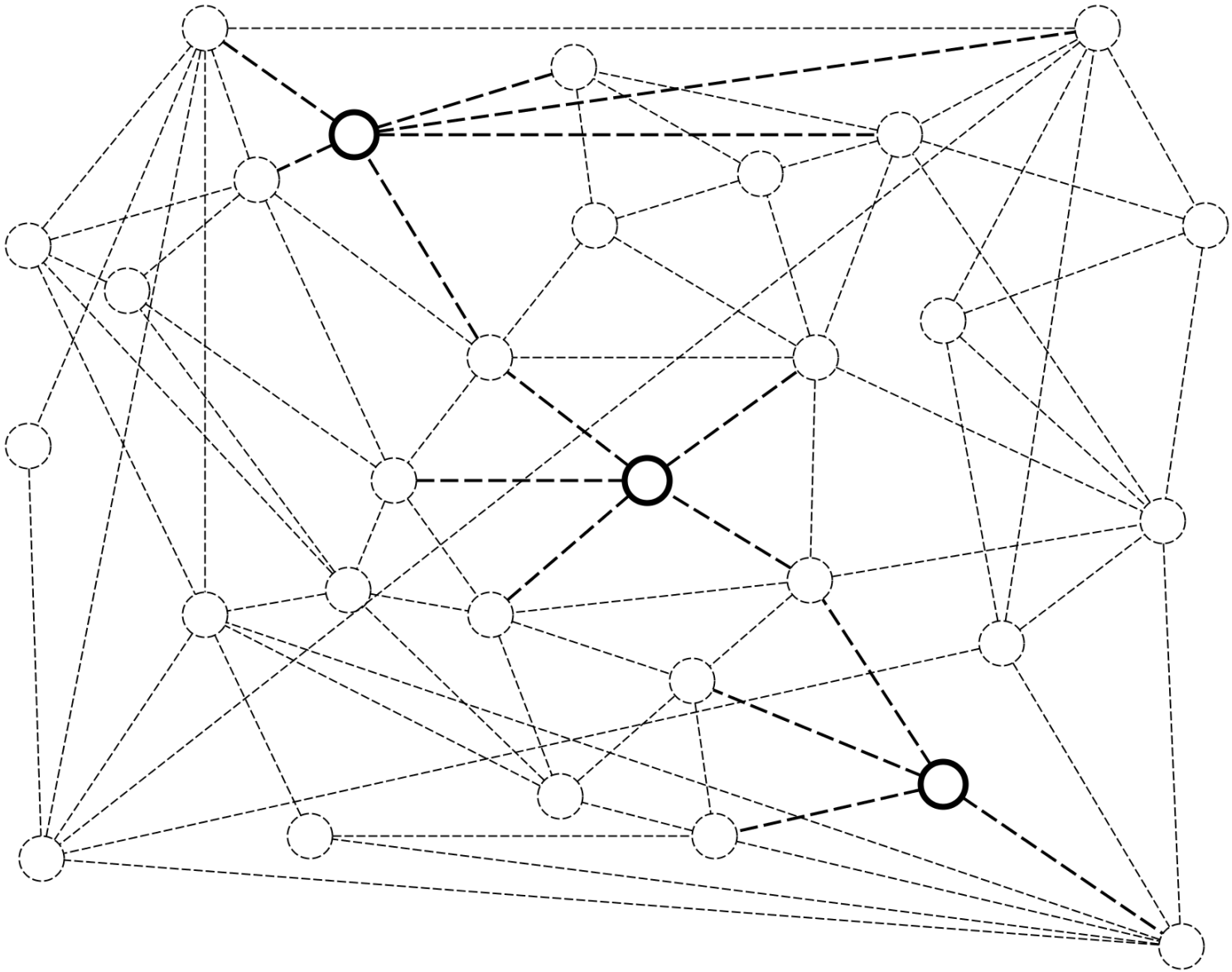
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 2



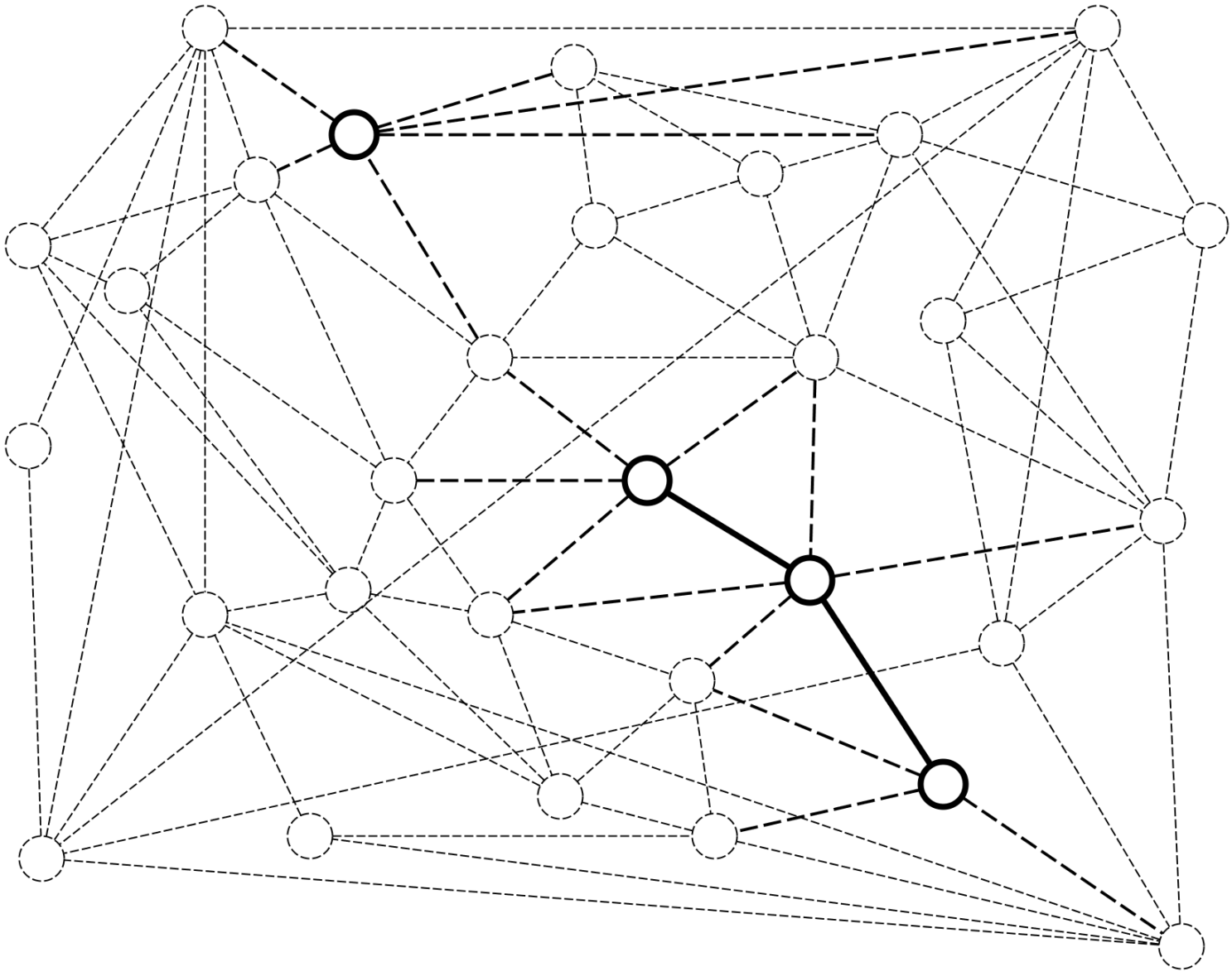
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 3



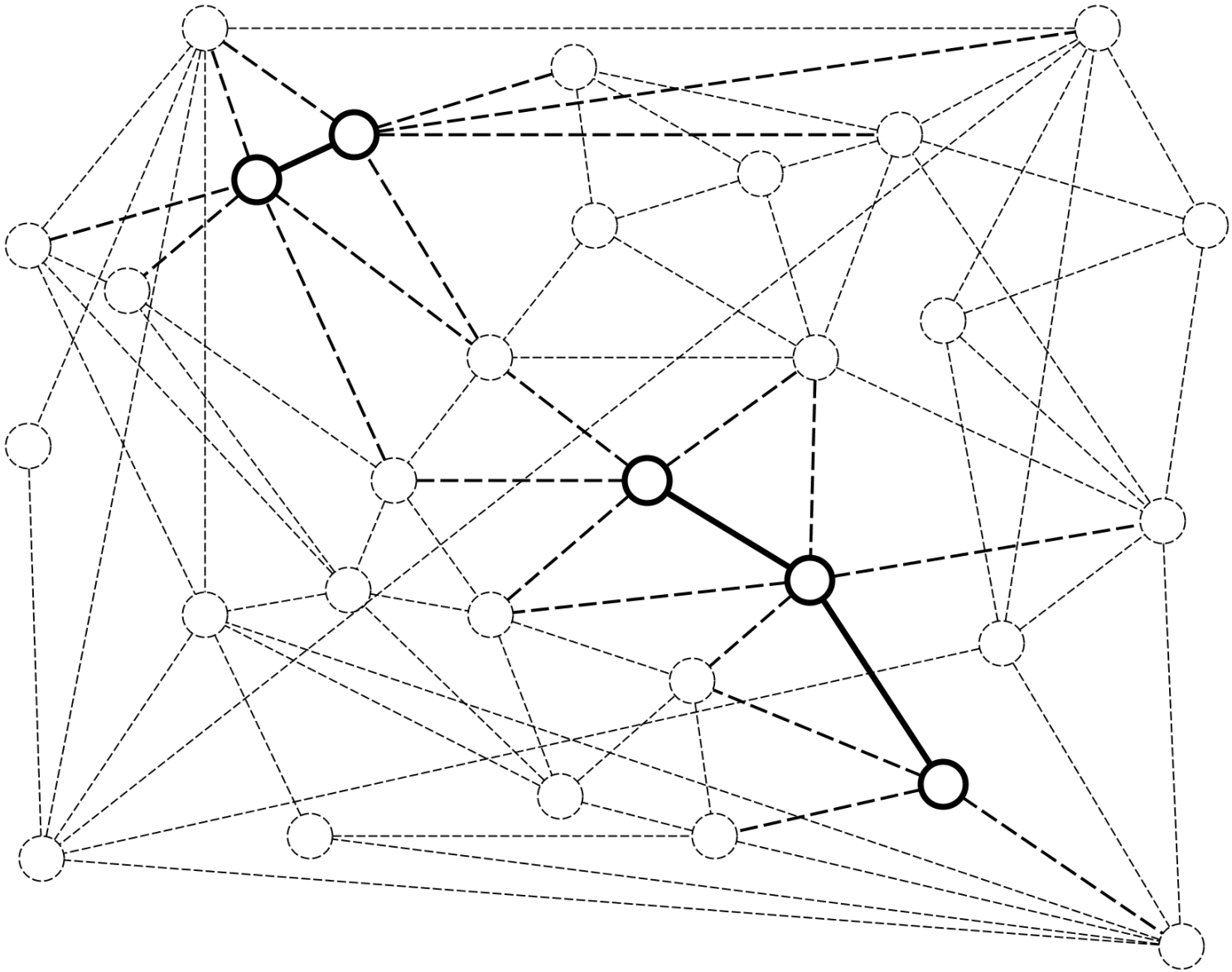
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 4



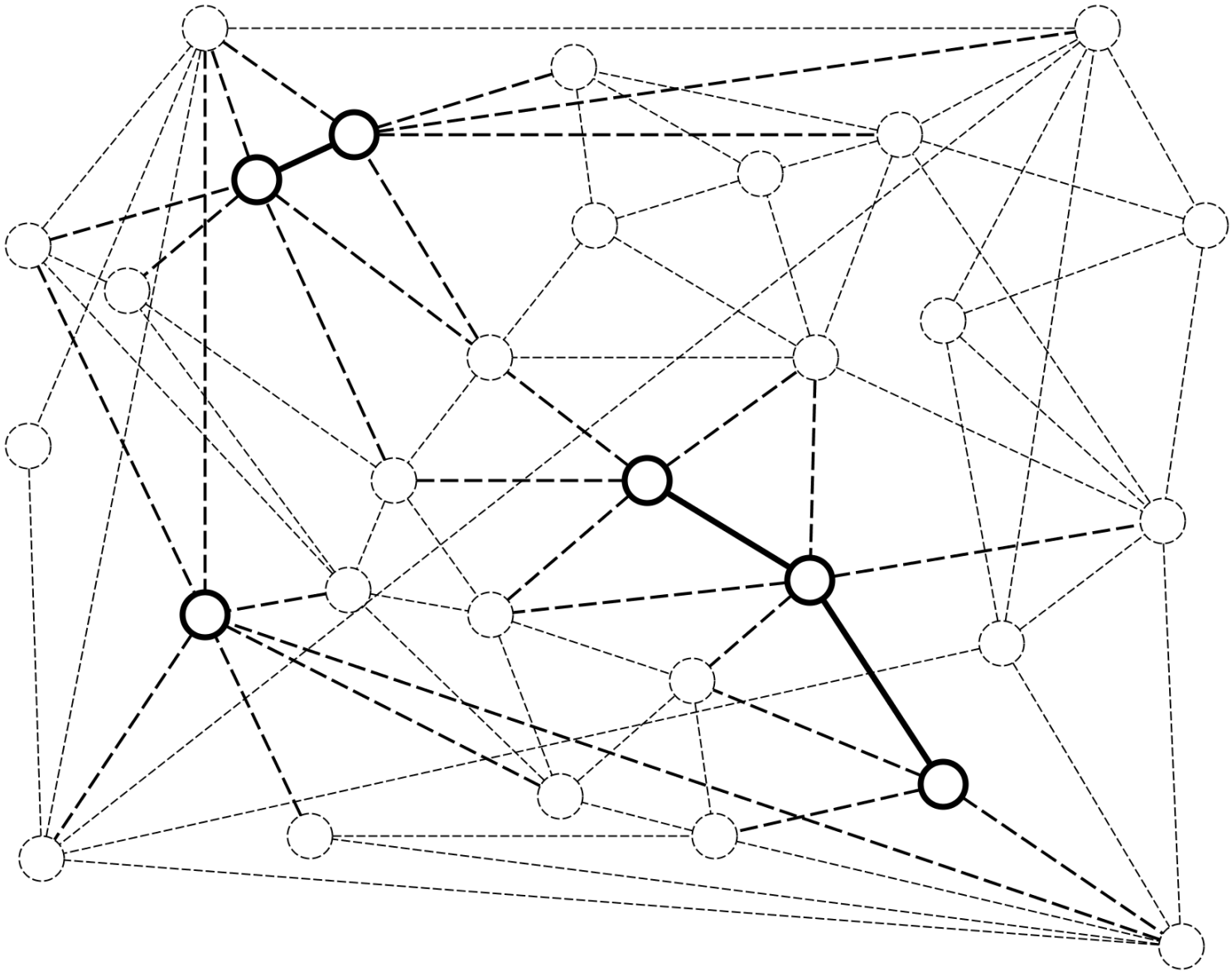
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 5



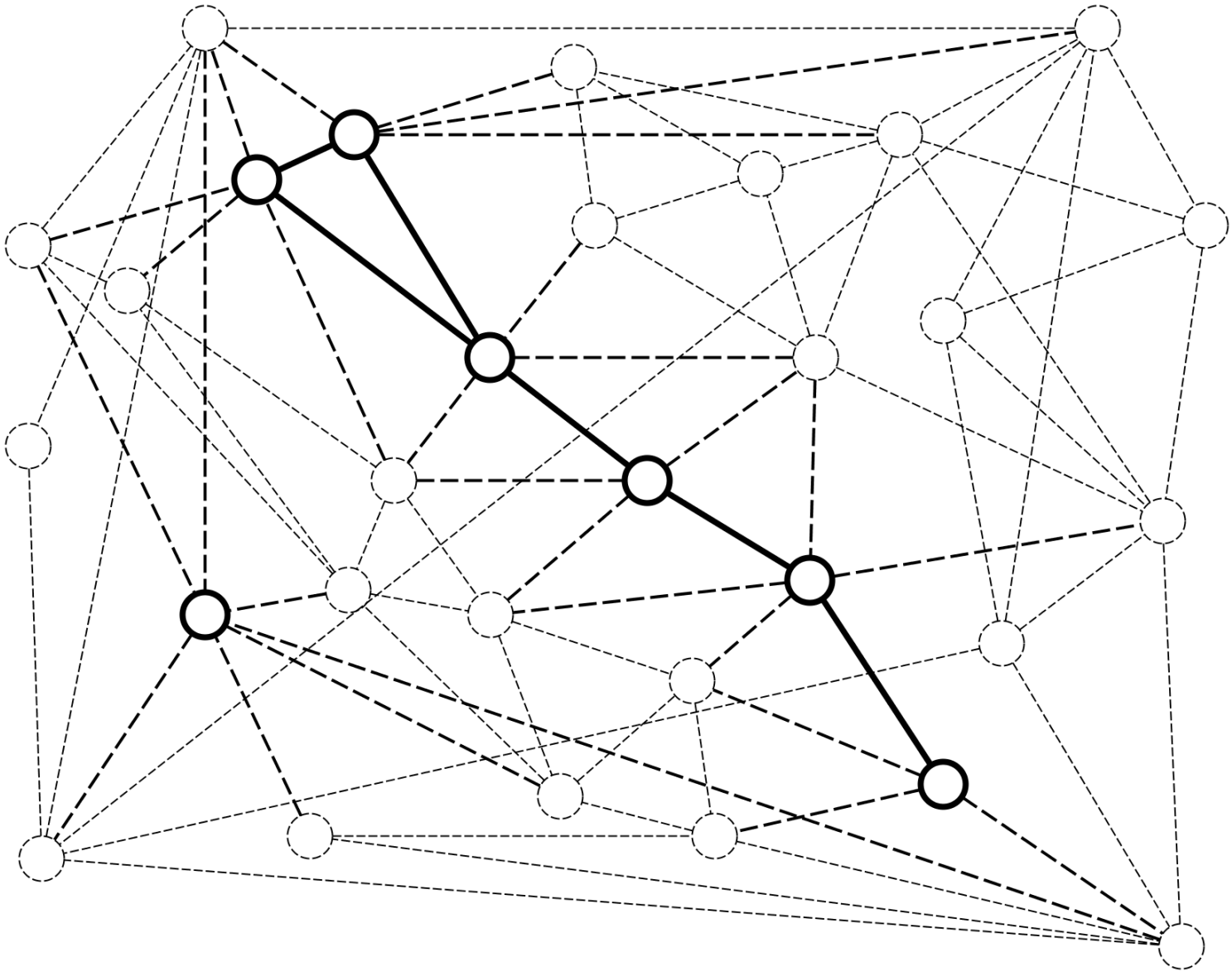
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 6



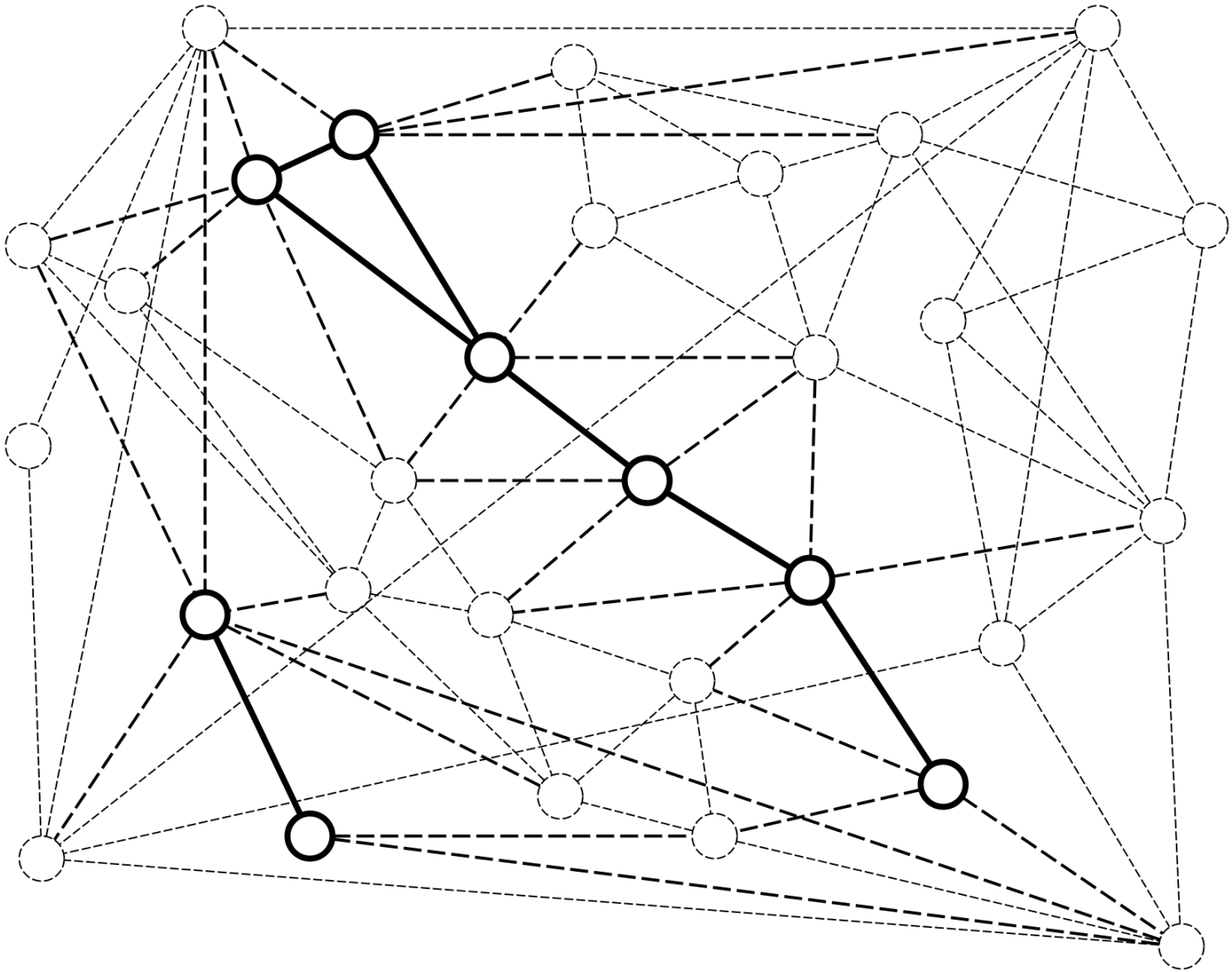
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 7



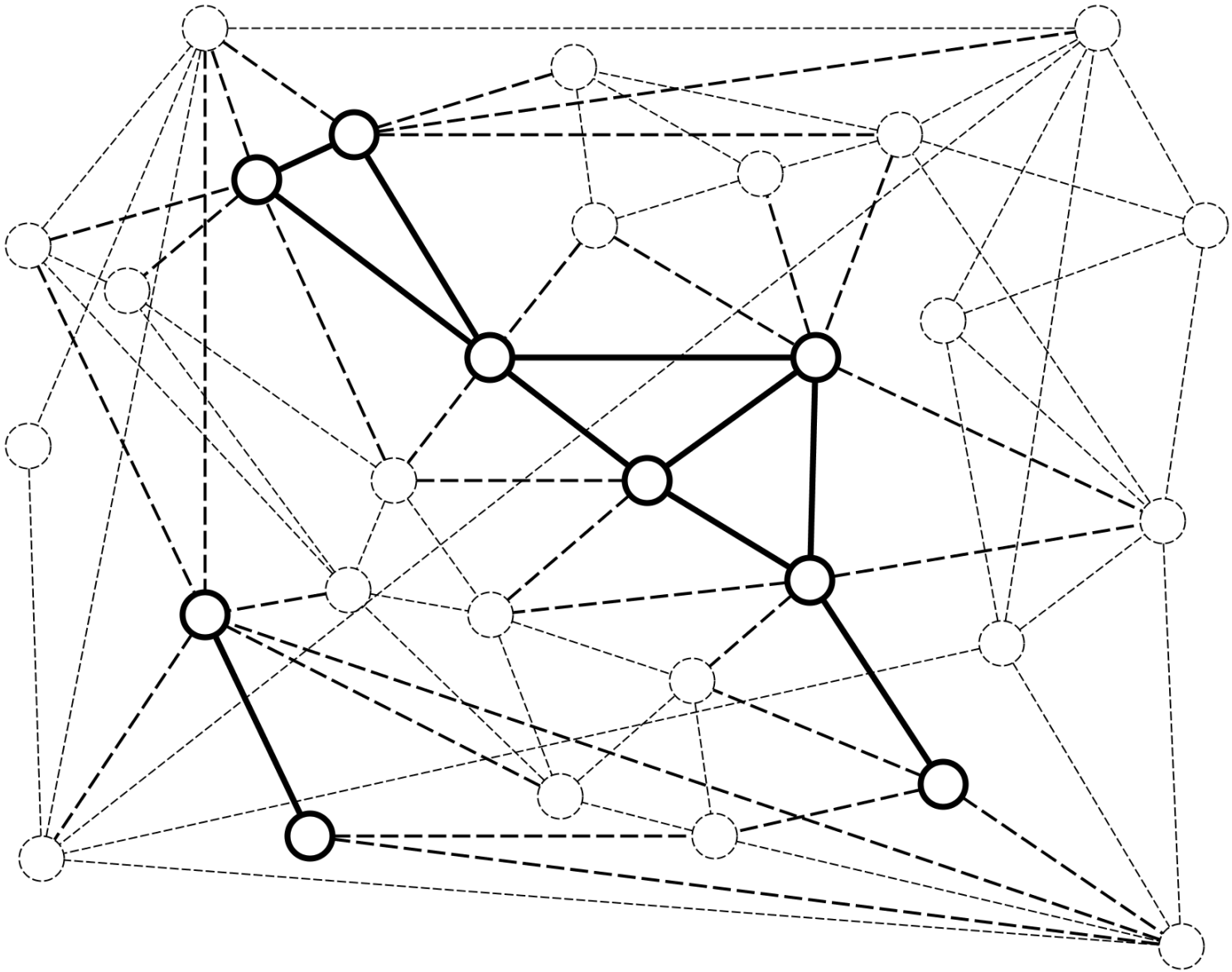
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 8



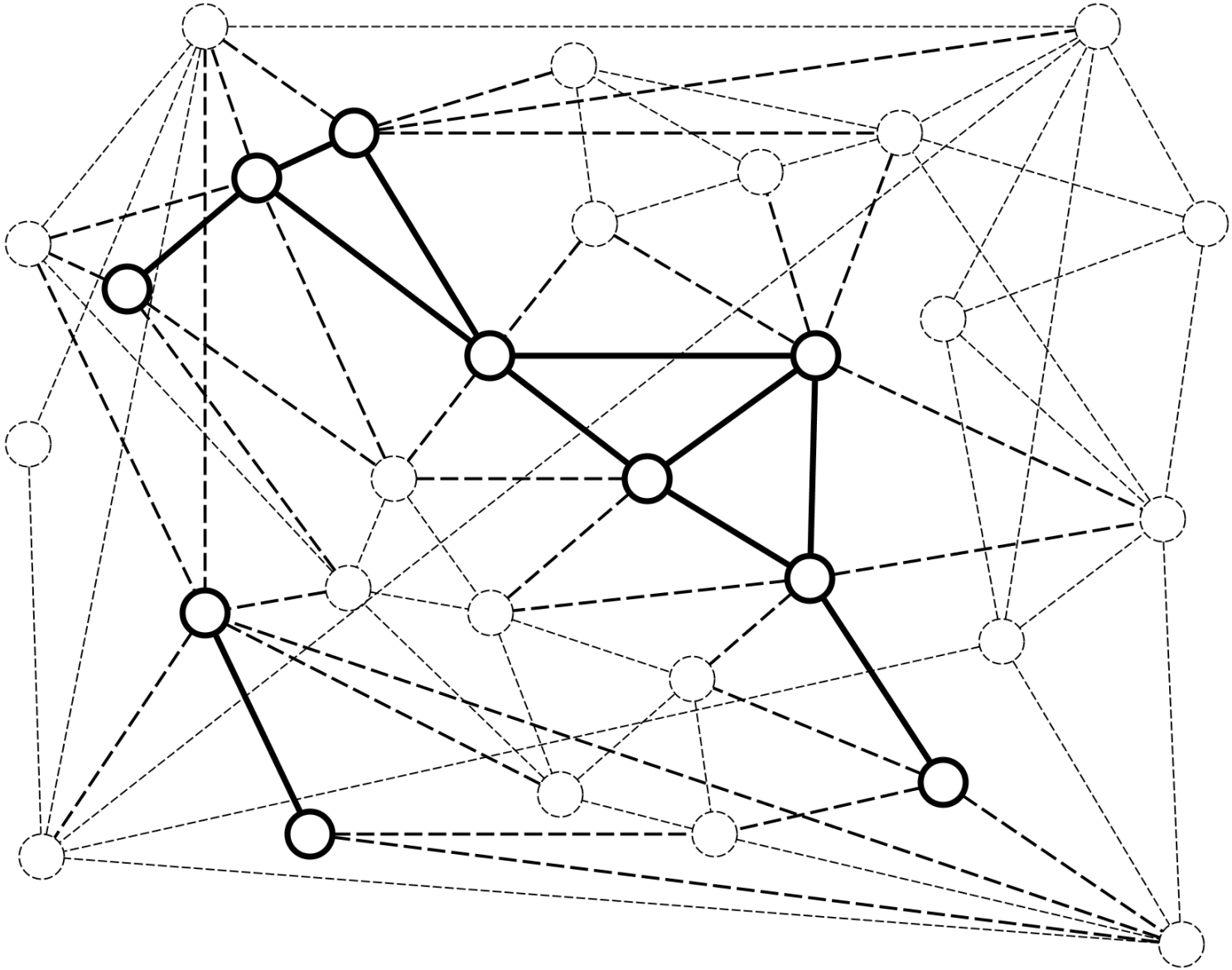
Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 9



Методы сэмплинга. Выбор случайной вершины

RN. Пример. Шаг 10



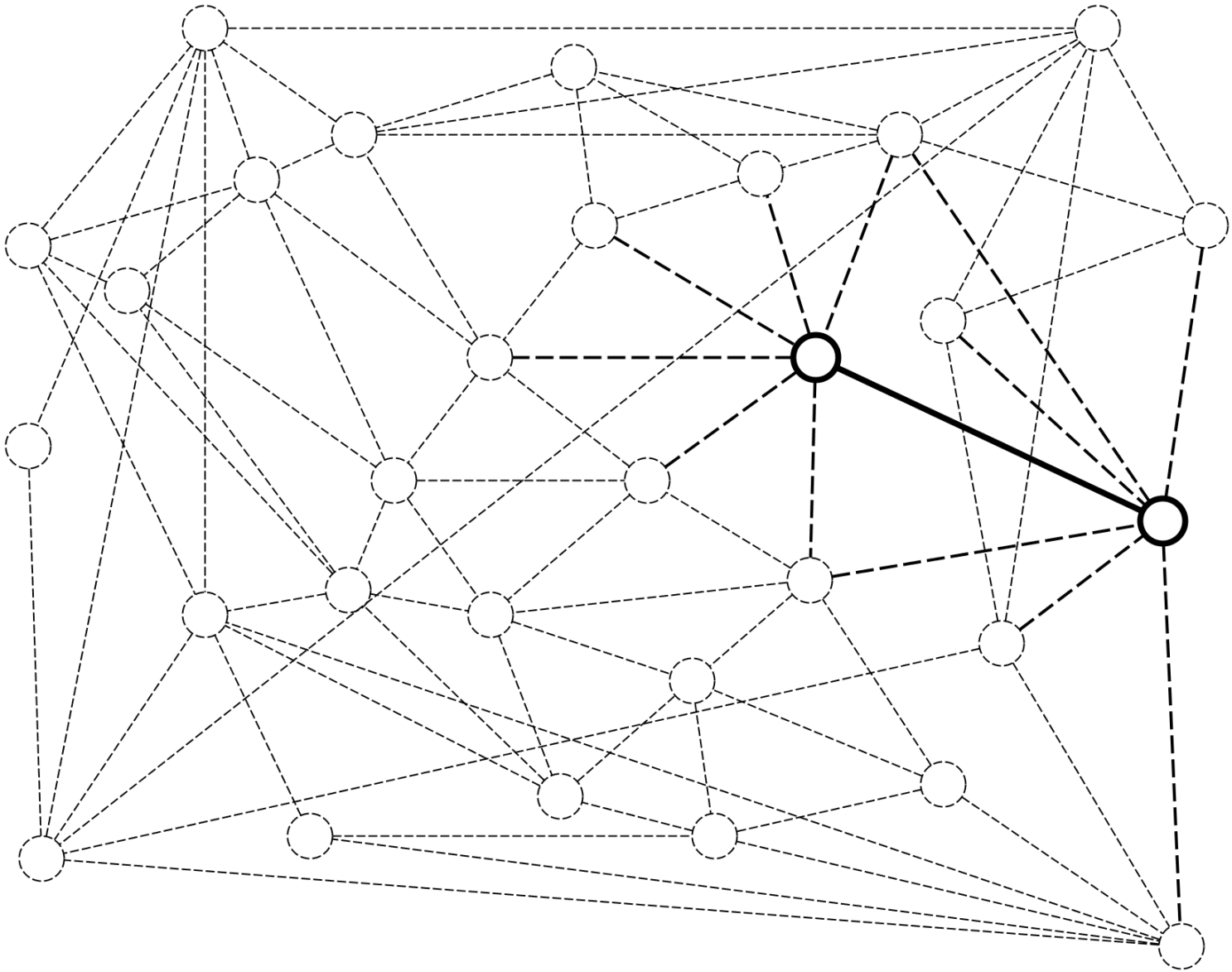
Методы сэмплинга. Выбор случайного ребра

Предполагается, что идентификаторы ребер графа известны

- Random edge (RE): Выбор случайного ребра
  - слабо связанный граф, большой диаметр
- Random node-edge (RNE):
  1. Выбор случайной вершины
  2. Выбор случайного ребра, инцидентной выбранной вершине
- Гибридный метод
  - С вероятностью p – RNE, с вероятностью $1 - p$ – RE

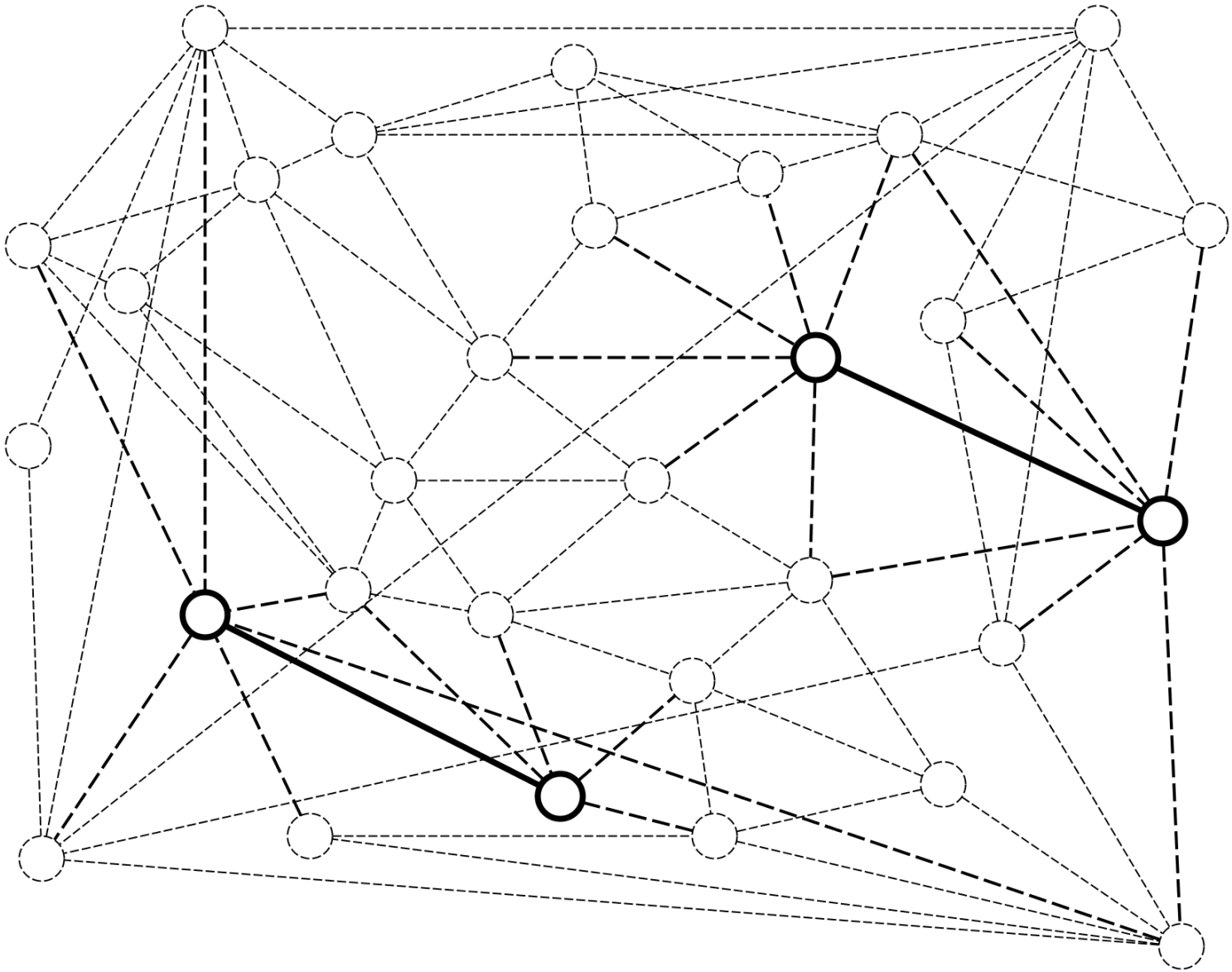
Методы сэмплинга. Выбор случайного ребра

RE. Пример. Шаг 1



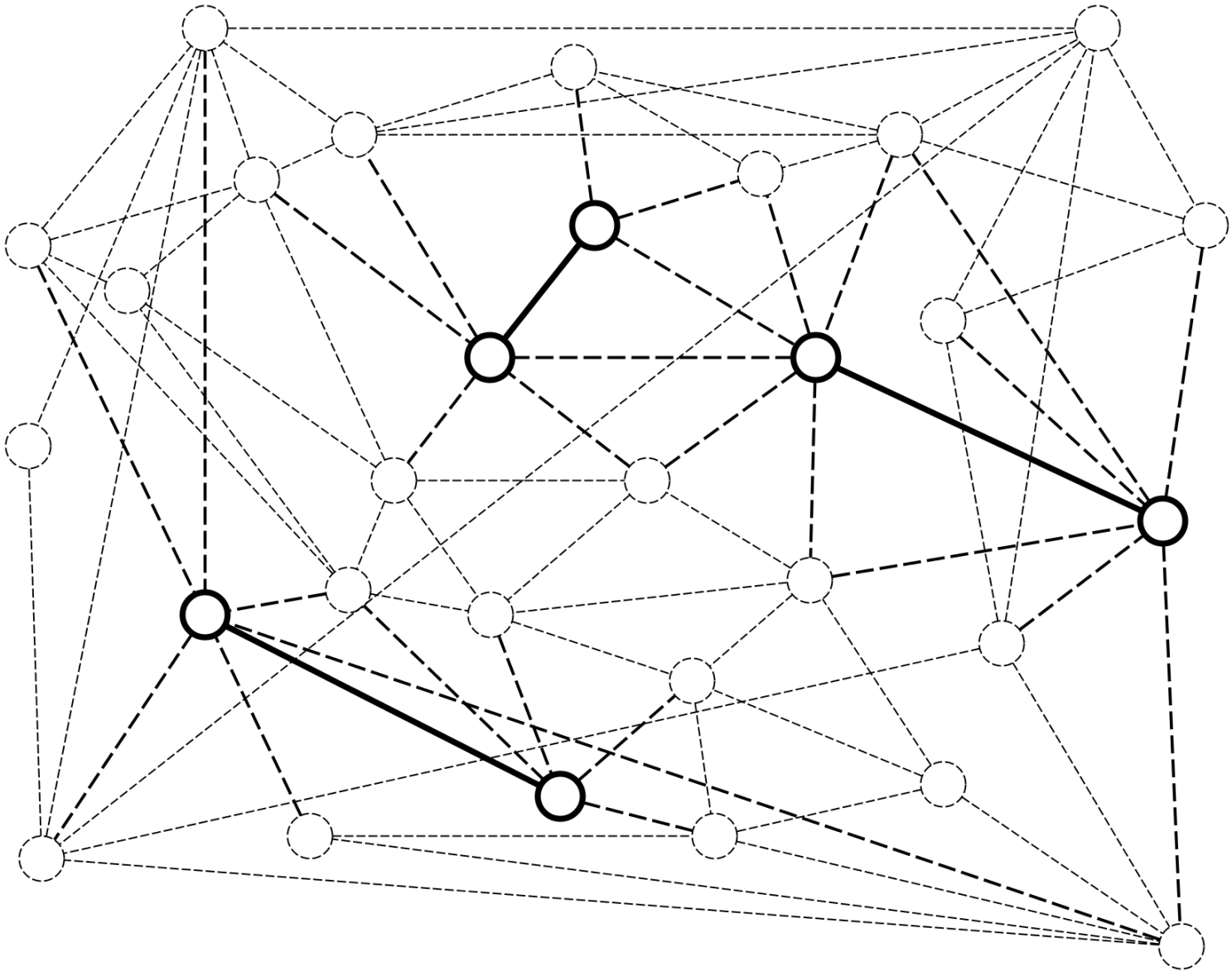
Методы сэмплинга. Выбор случайного ребра

RE. Пример. Шаг 2



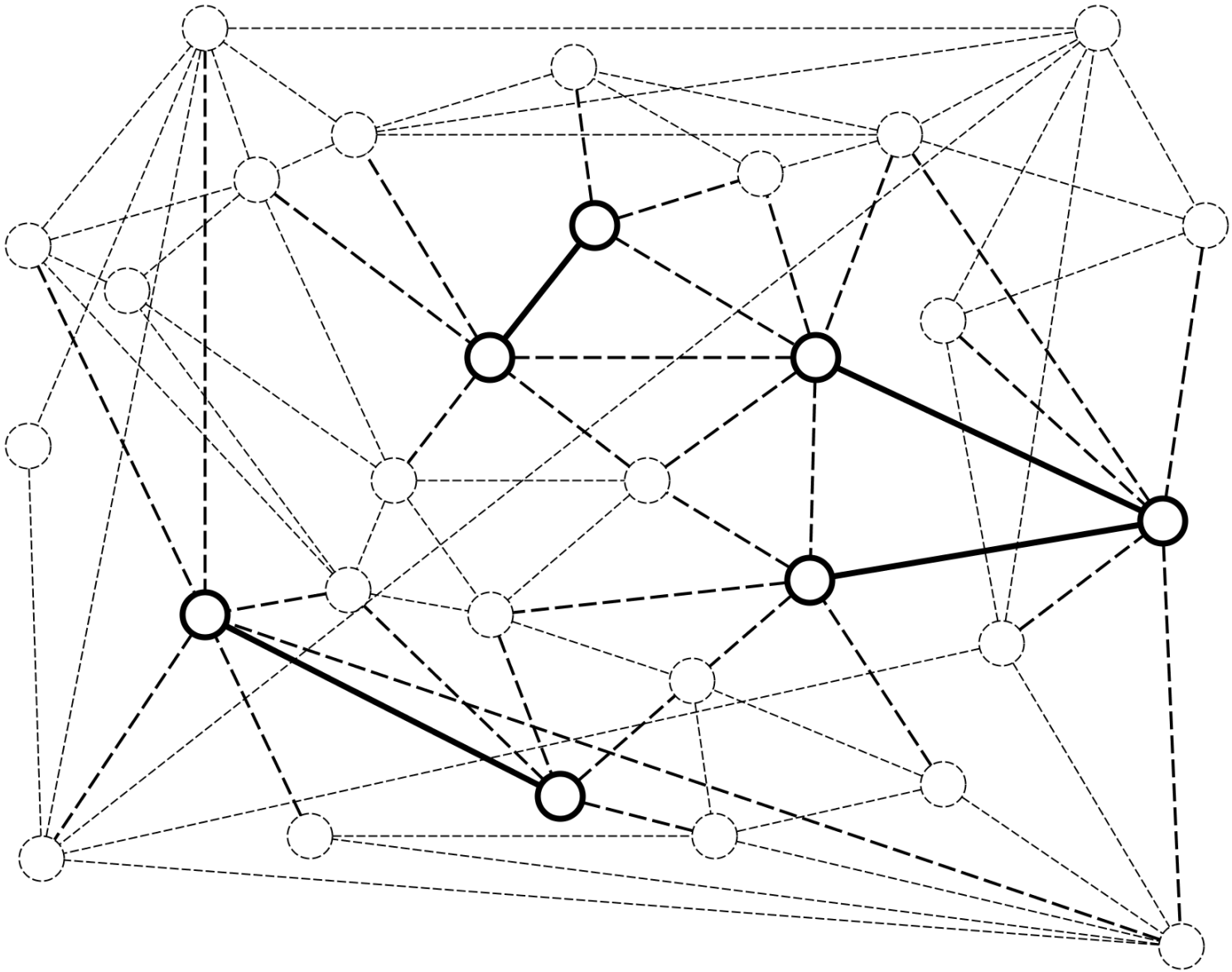
Методы сэмплинга. Выбор случайного ребра

RE. Пример. Шаг 3



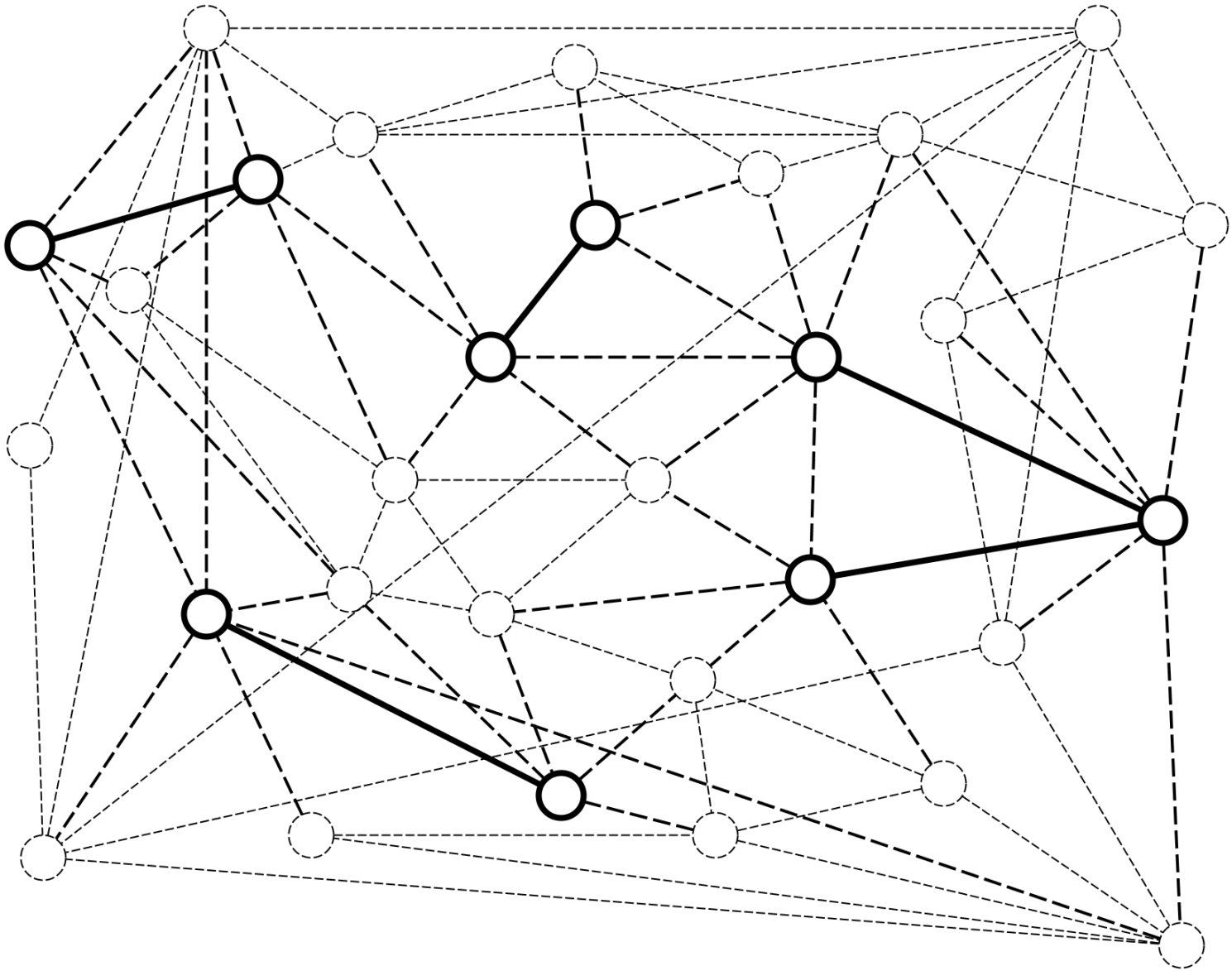
Методы сэмплинга. Выбор случайного ребра

RE. Пример. Шаг 4



Методы сэмплинга. Выбор случайного ребра

RE. Пример. Шаг 5



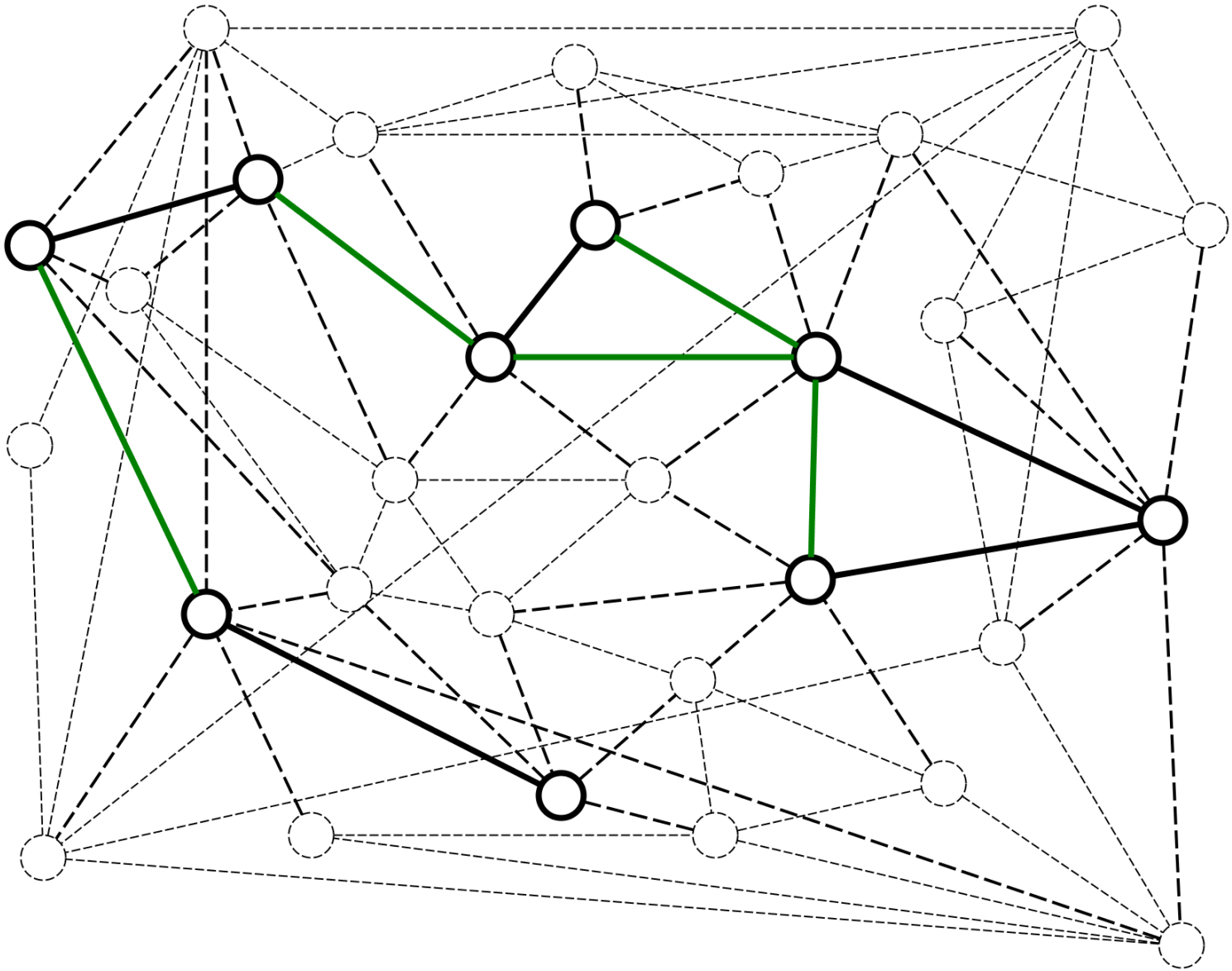
Методы сэмплинга. Выбор случайного ребра

Предполагается, что идентификаторы ребер (или вершин) графа известны

- Induced Edge Sampling:
  1. Несколько раз равномерно выбирать ребра (и соответствующие узлы)
  2. Добавить в сэмпл ребра, существующие между выбранными вершинами

Методы сэмплинга. Выбор случайного ребра

Included Edge. Пример

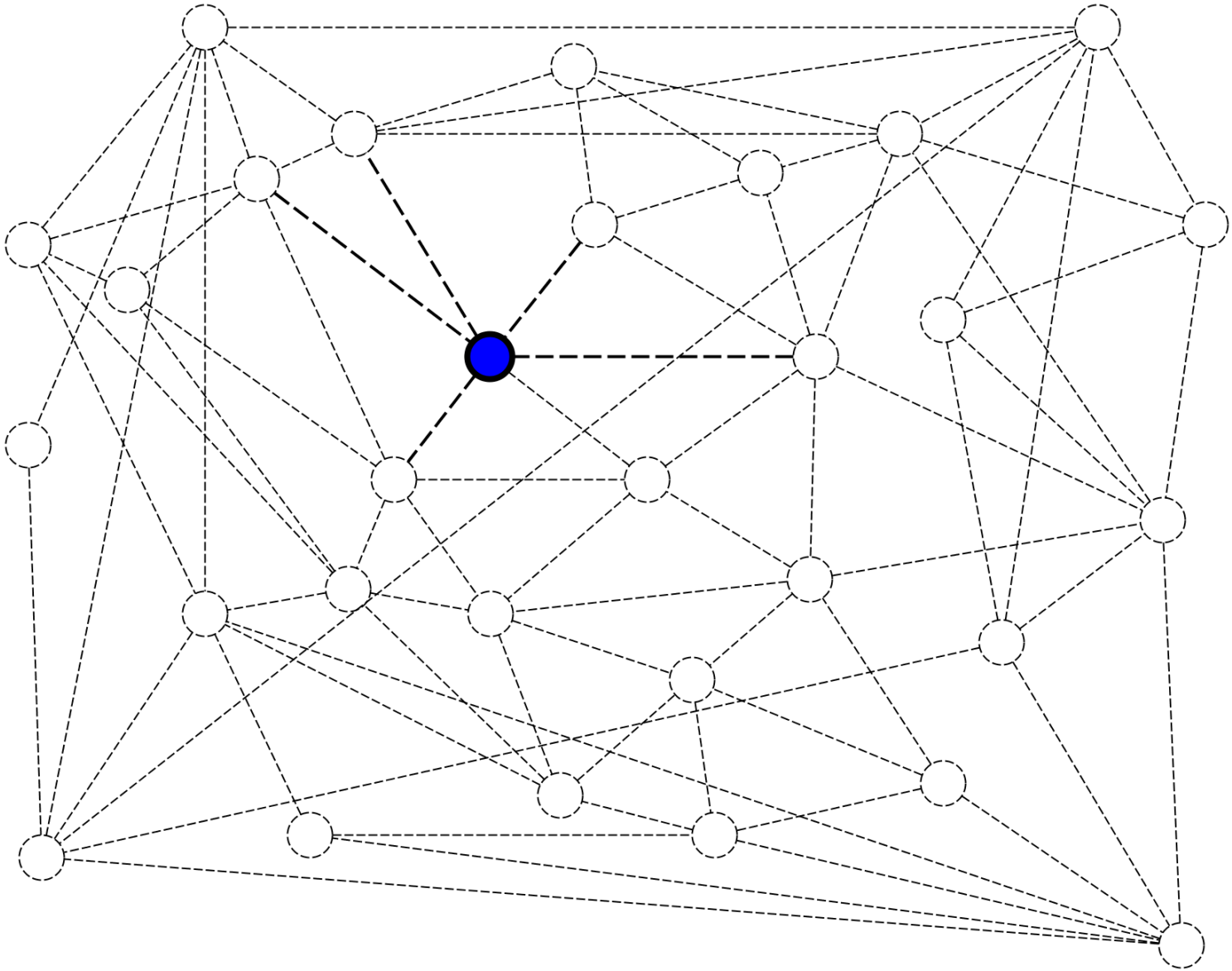


Методы сэмплинга. Исследование социальной сети

- Случайные блуждания:
  - Random Walk (RW):
    1. Случайный выбор инцидентного ребра
    2. Переход к соответствующей вершине

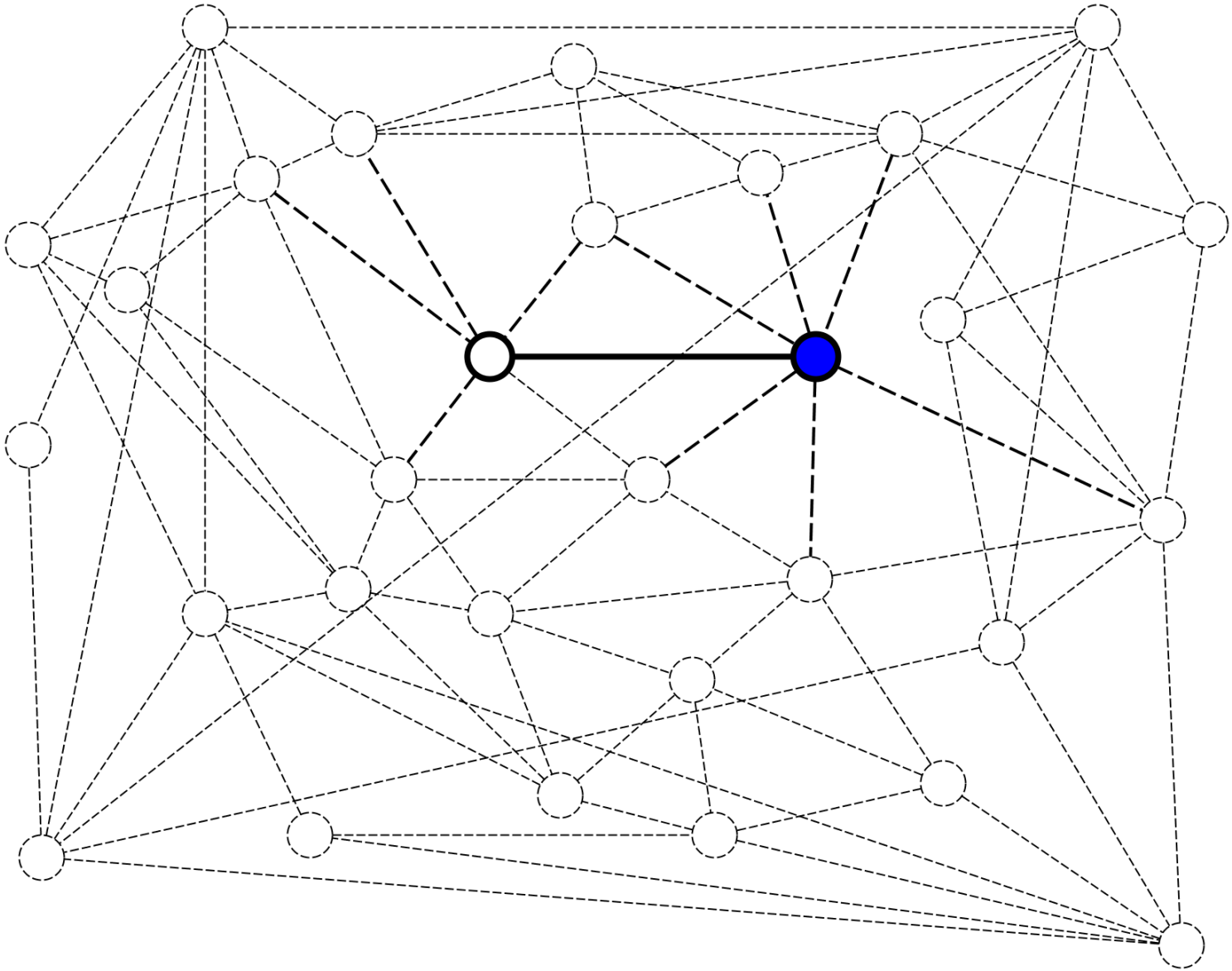
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 1



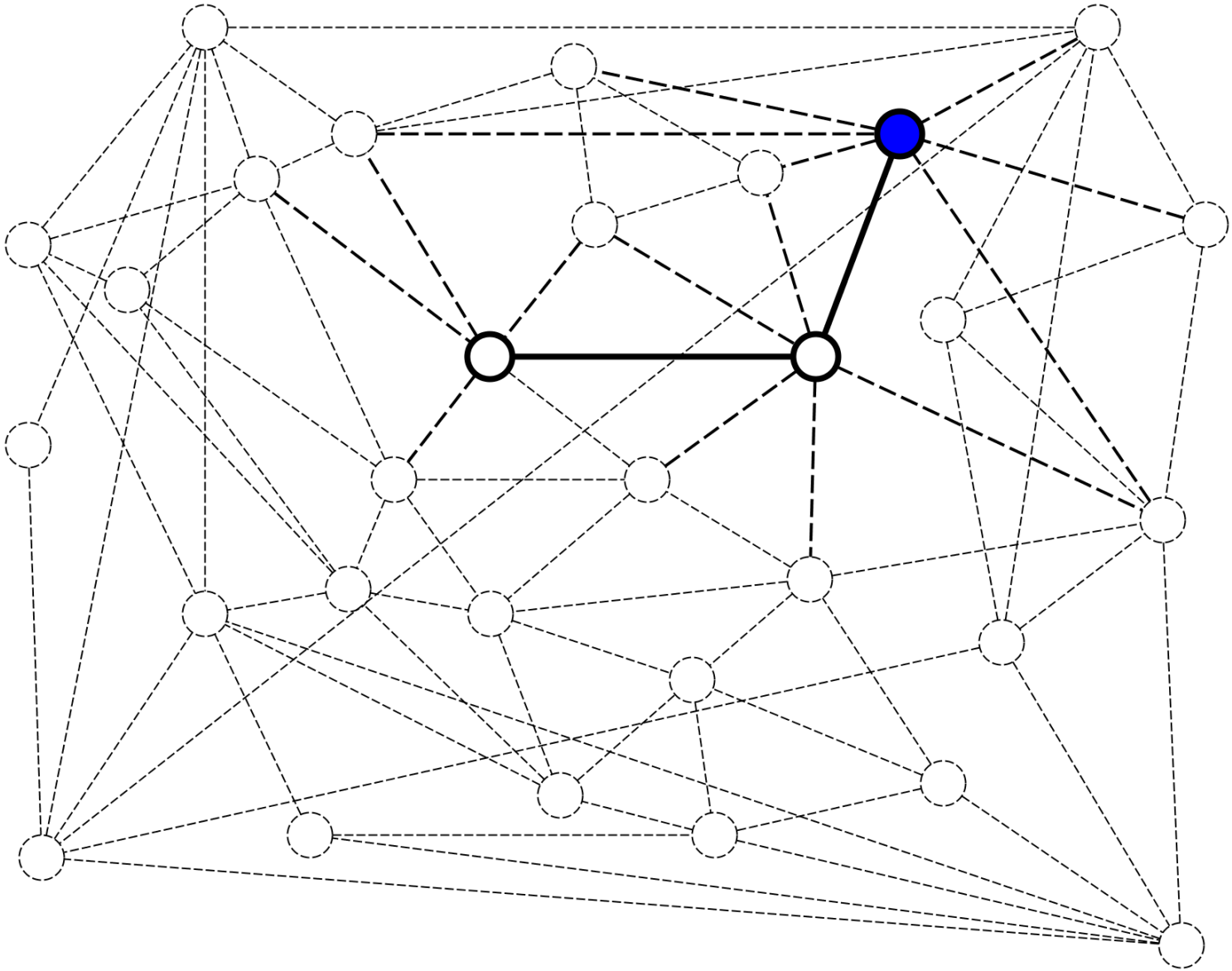
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг2



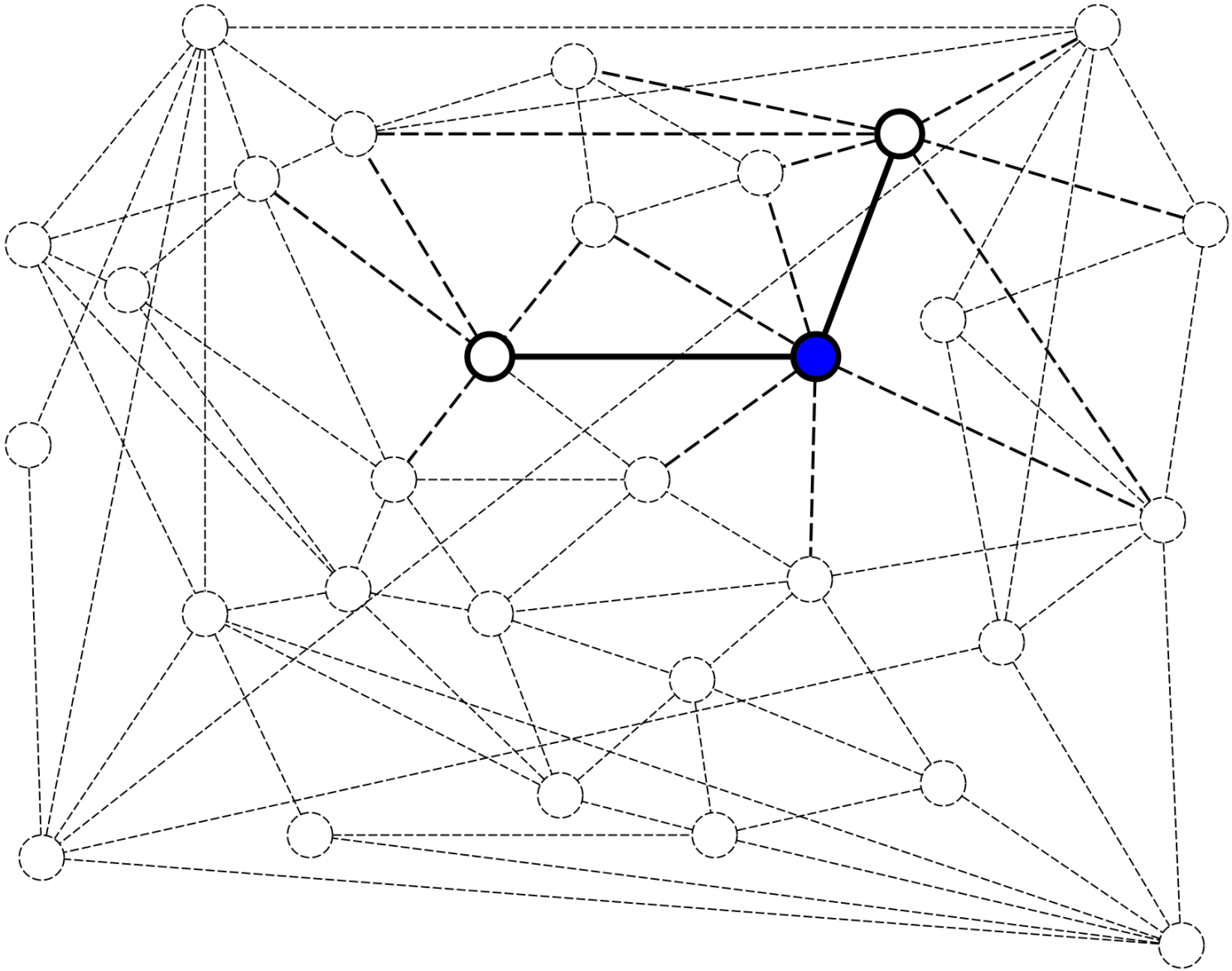
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 3



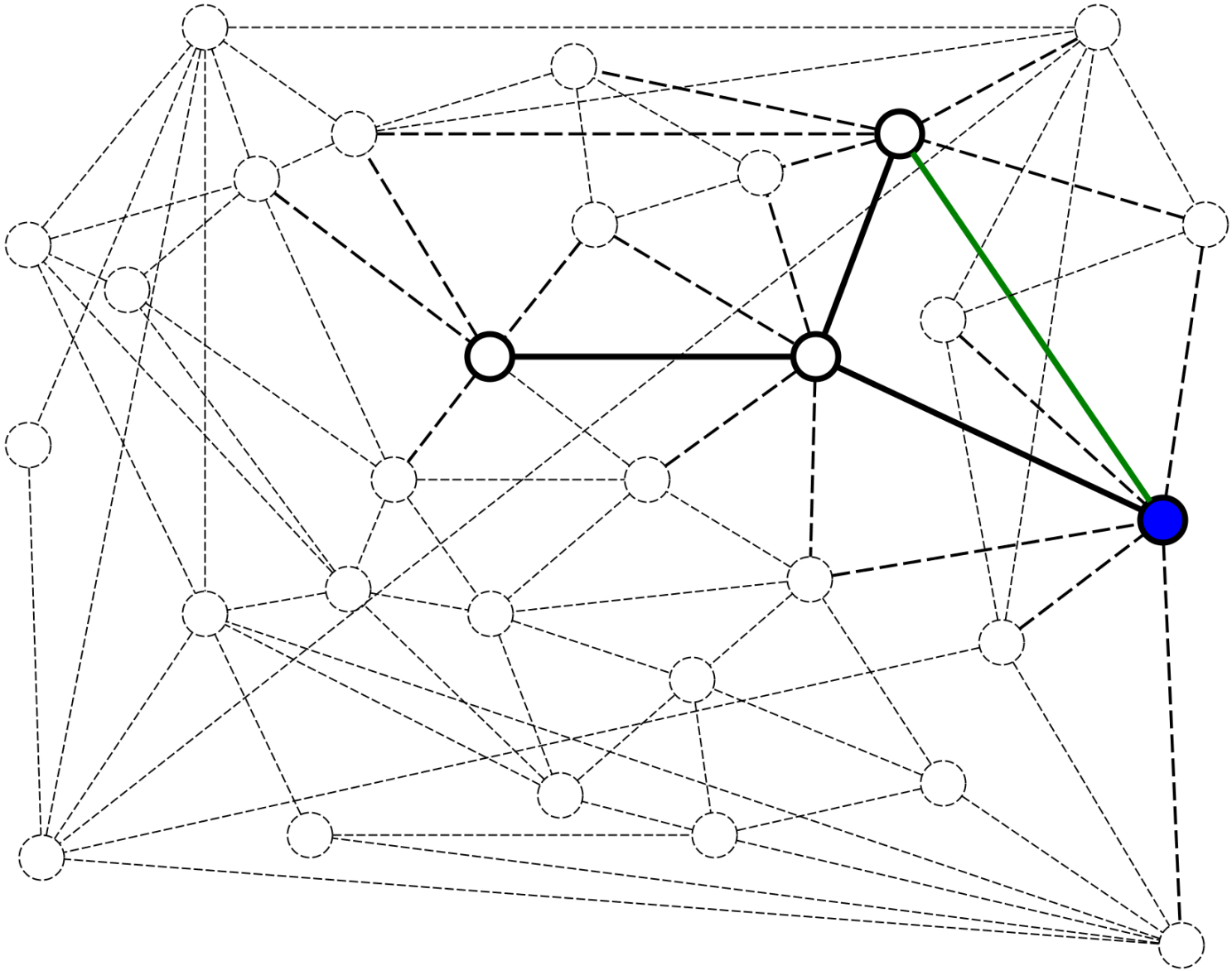
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 4



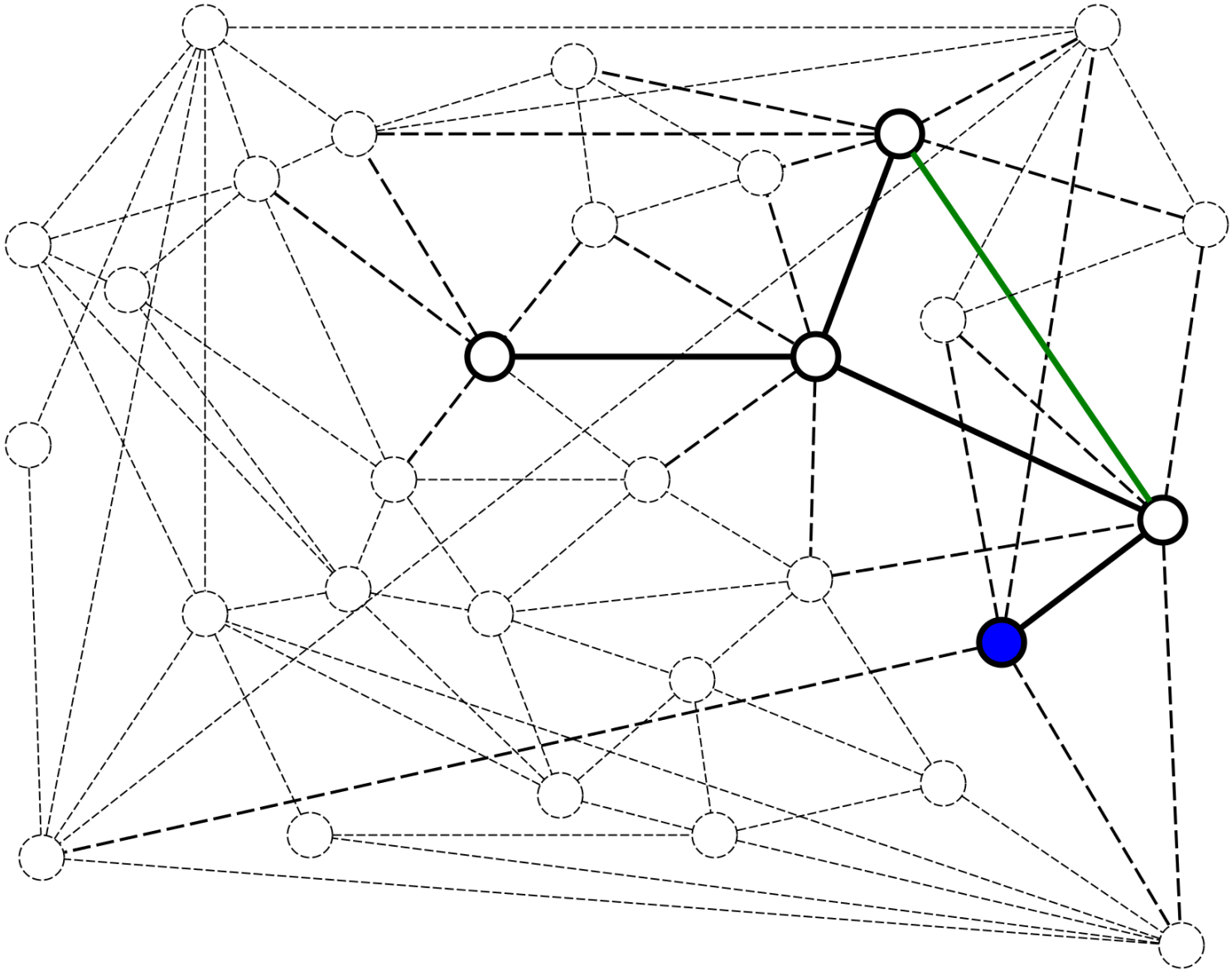
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 5



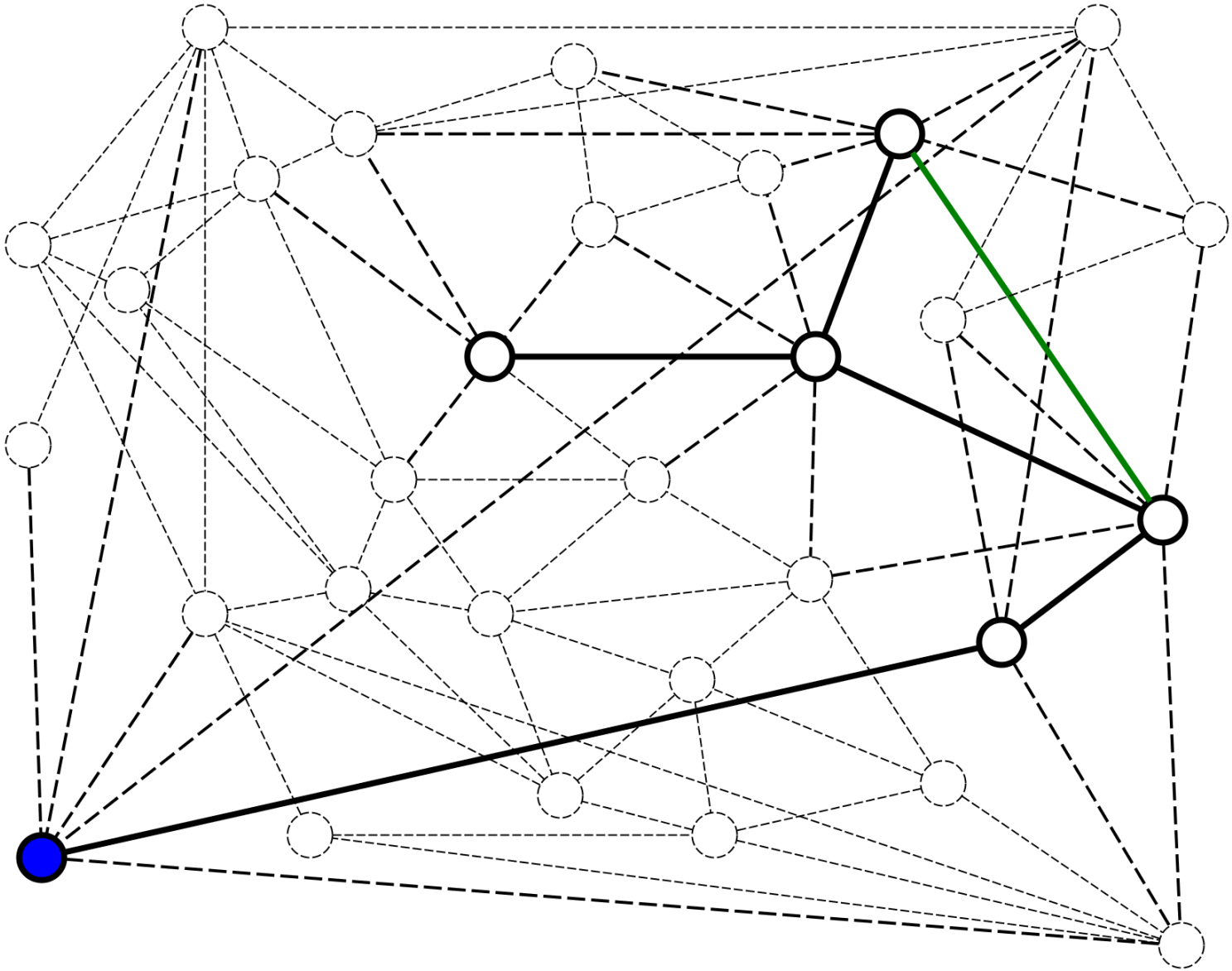
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 6



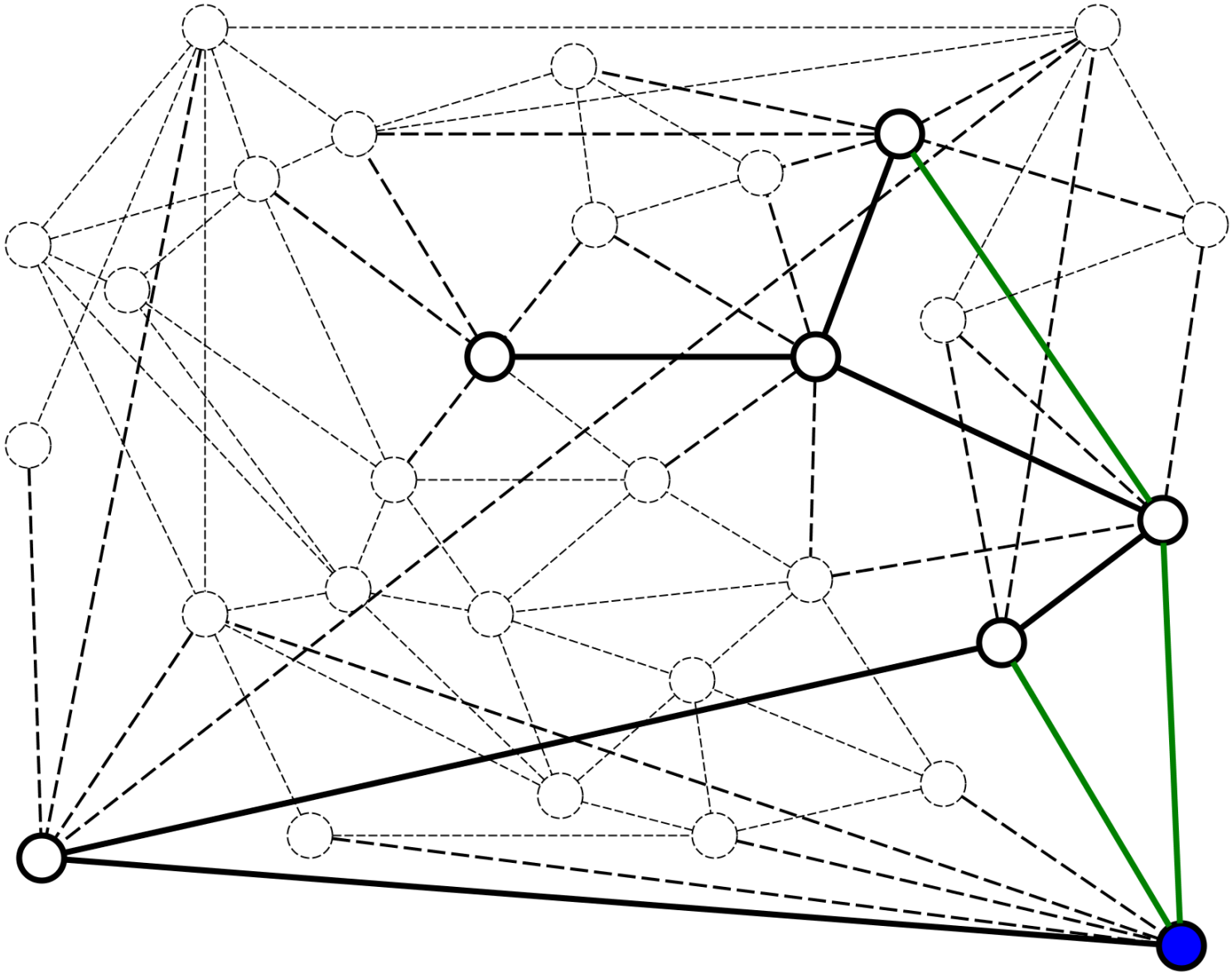
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 7



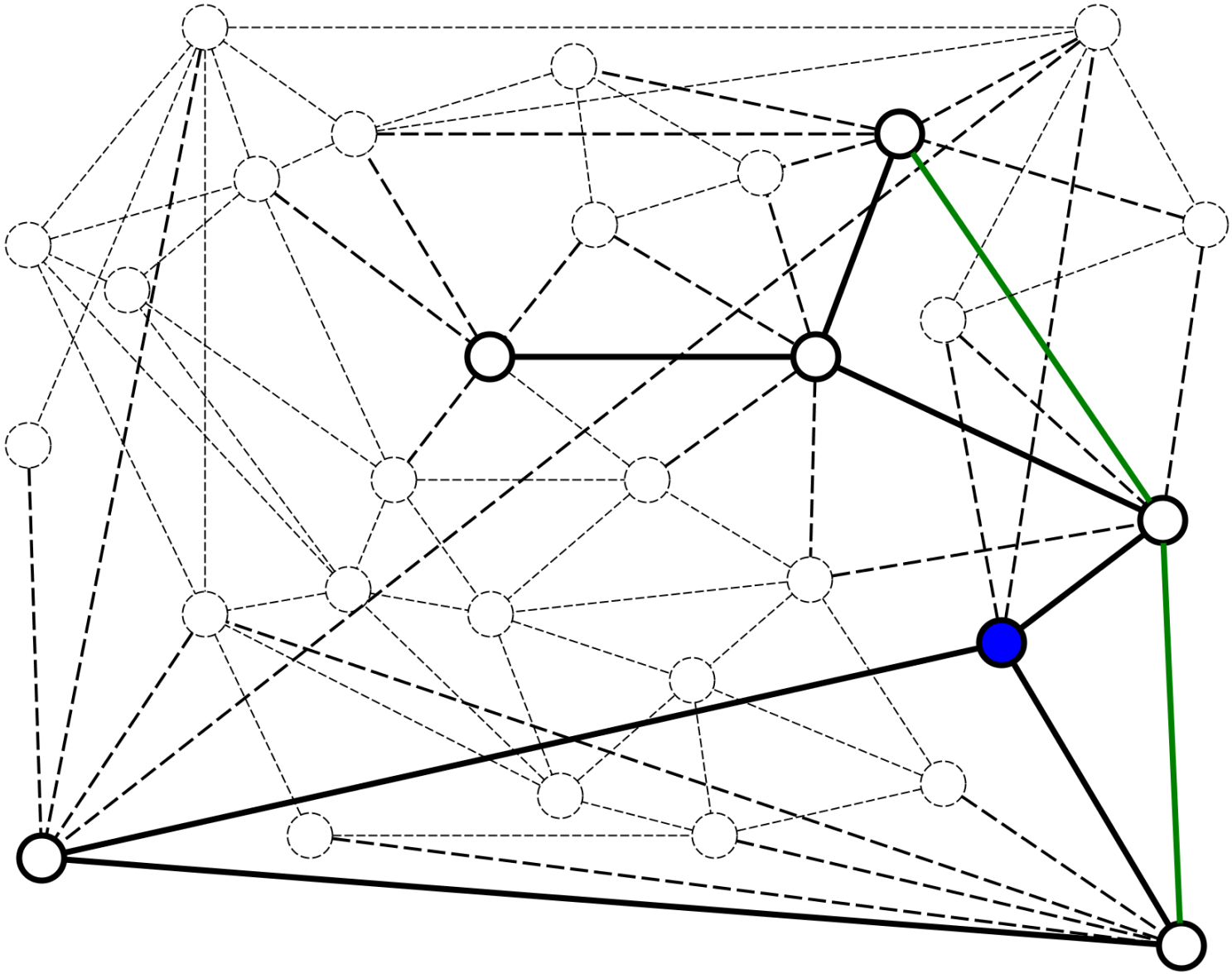
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 8



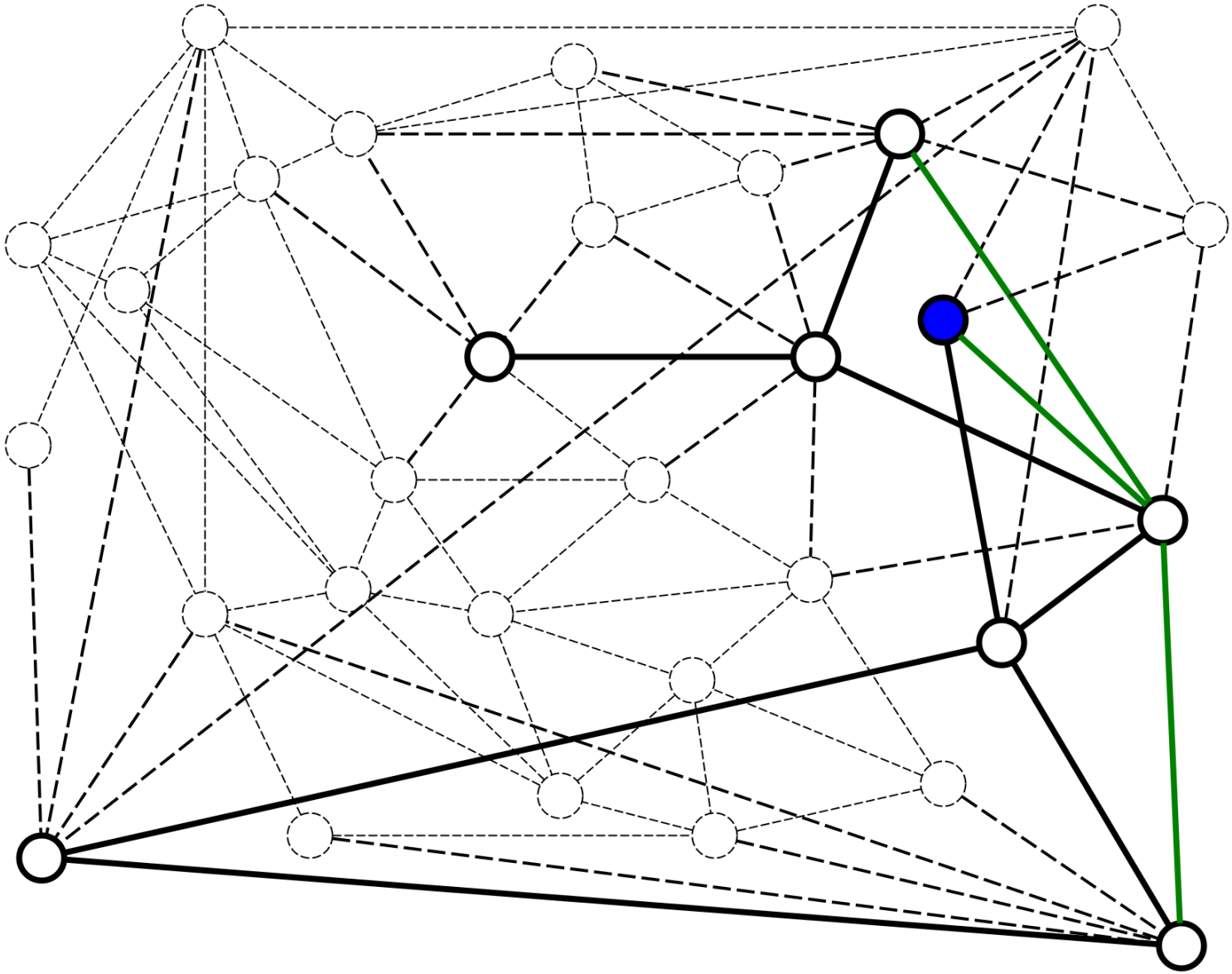
Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 9



Методы сэмплинга. Исследование социальной сети

Случайные блуждания (RW). Пример. Шаг 10

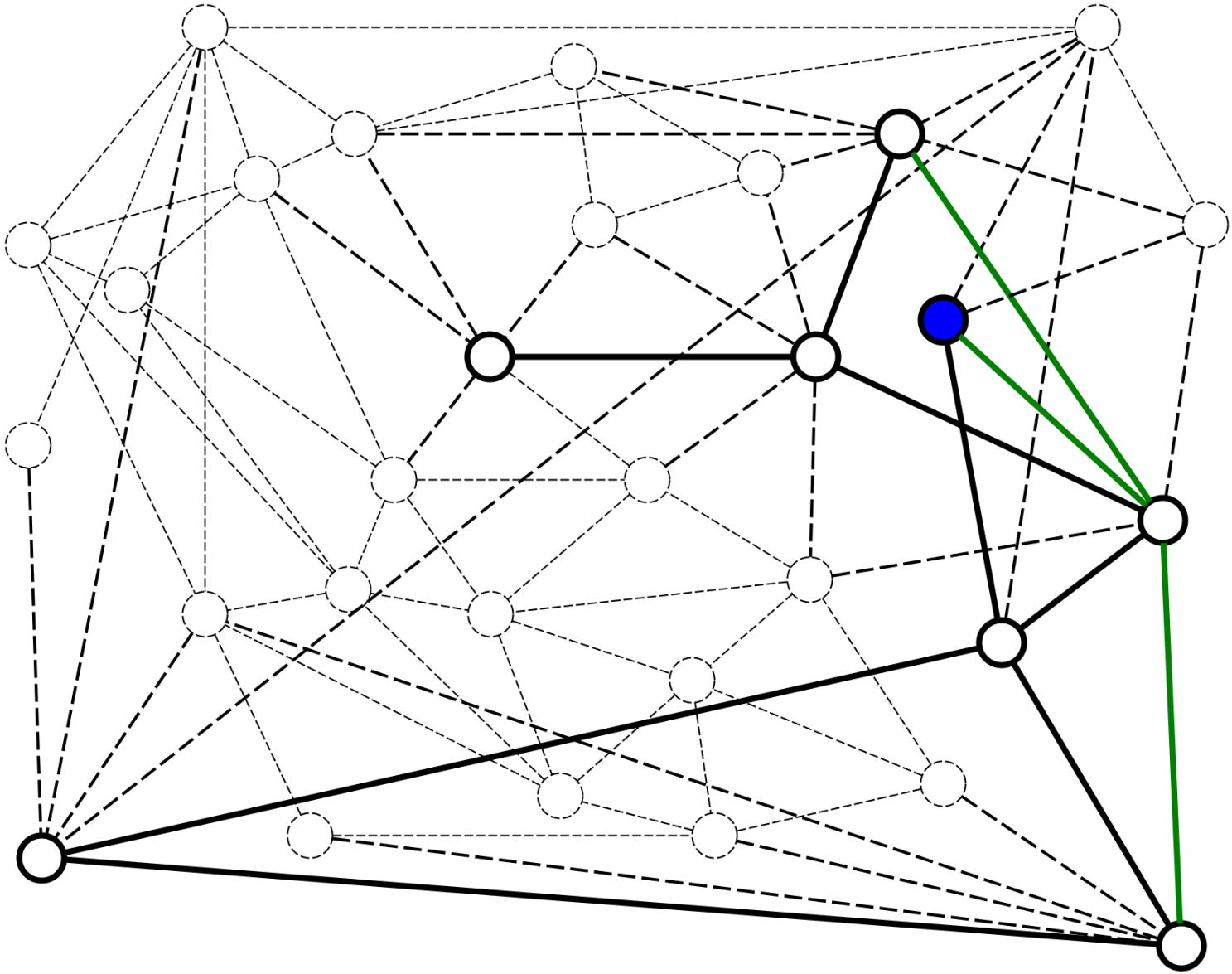


Методы сэмплинга. Исследование социальной сети

- Случайные блуждания:
  - Random Walk (RW):
    1. Случайный выбор инцидентного ребра
    2. Переход к соответствующей вершине
  - Random Walk with Restart (RWR):
    - ★ Модификация: переход в начальную вершину с вероятностью p

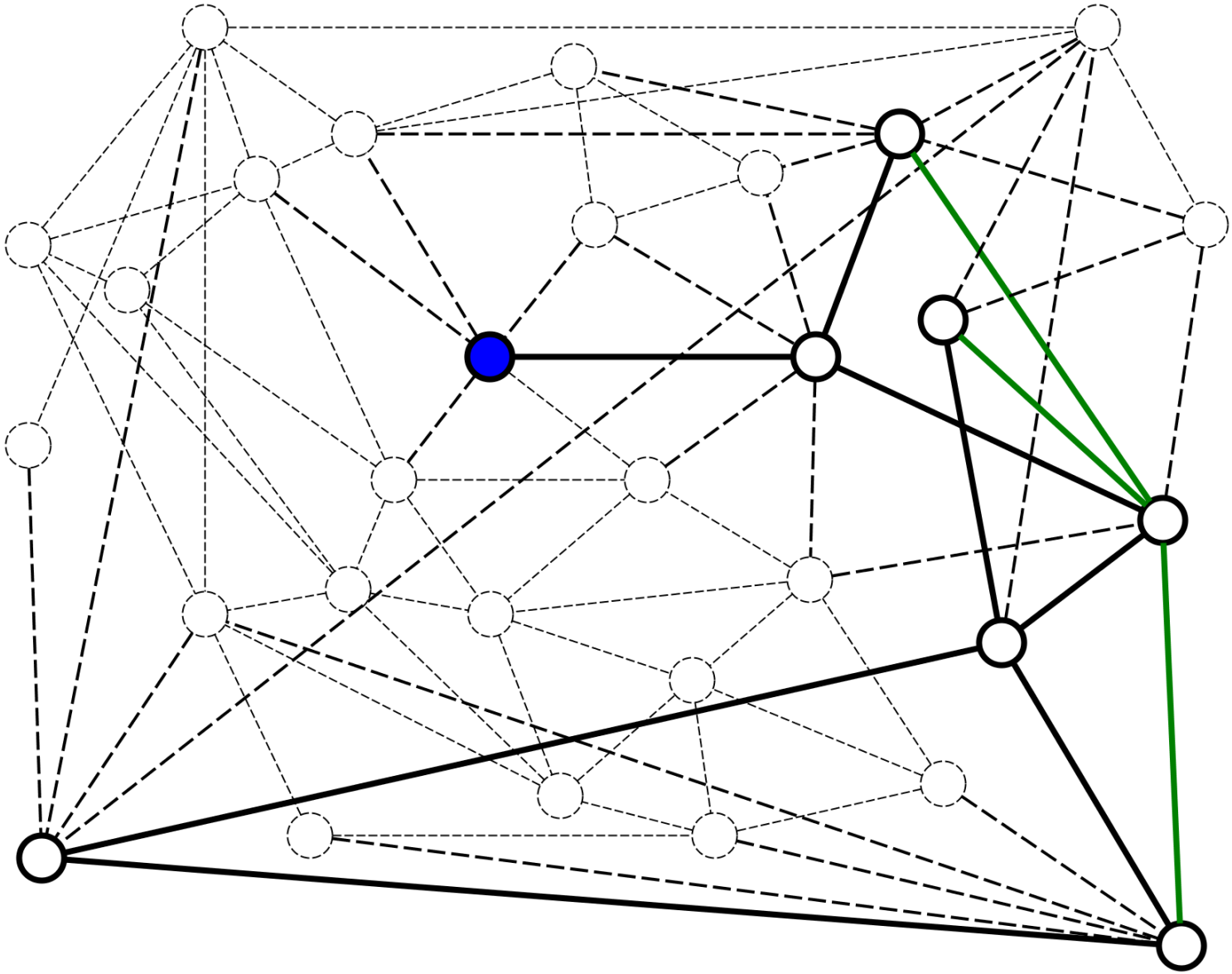
Методы сэмплинга. Исследование социальной сети

Случайные блуждания с рестартом (RWR). Пример



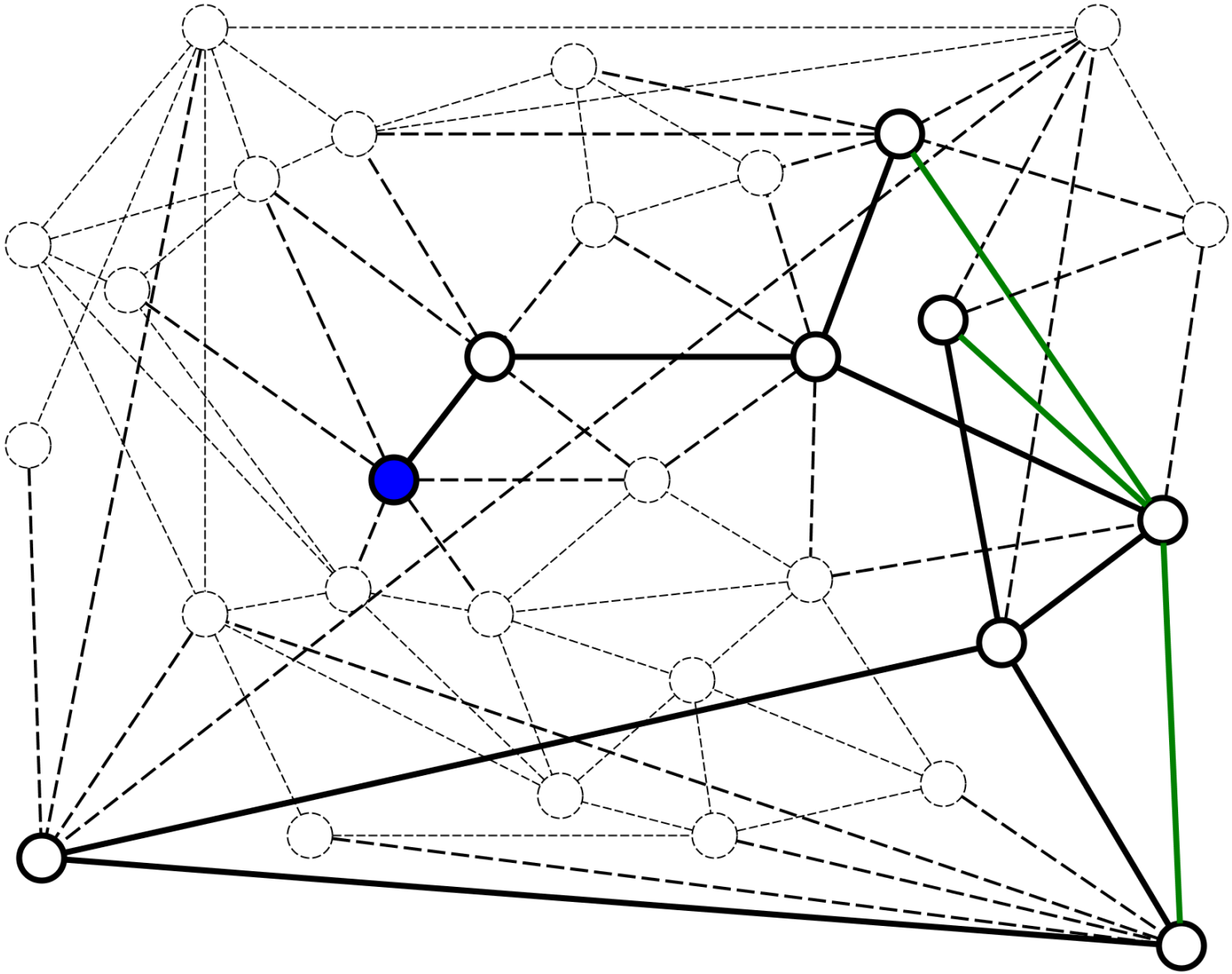
Методы сэмплинга. Исследование социальной сети

Случайные блуждания с рестартом (RWR). Пример



Методы сэмплинга. Исследование социальной сети

Случайные блуждания с рестартом (RWR). Пример

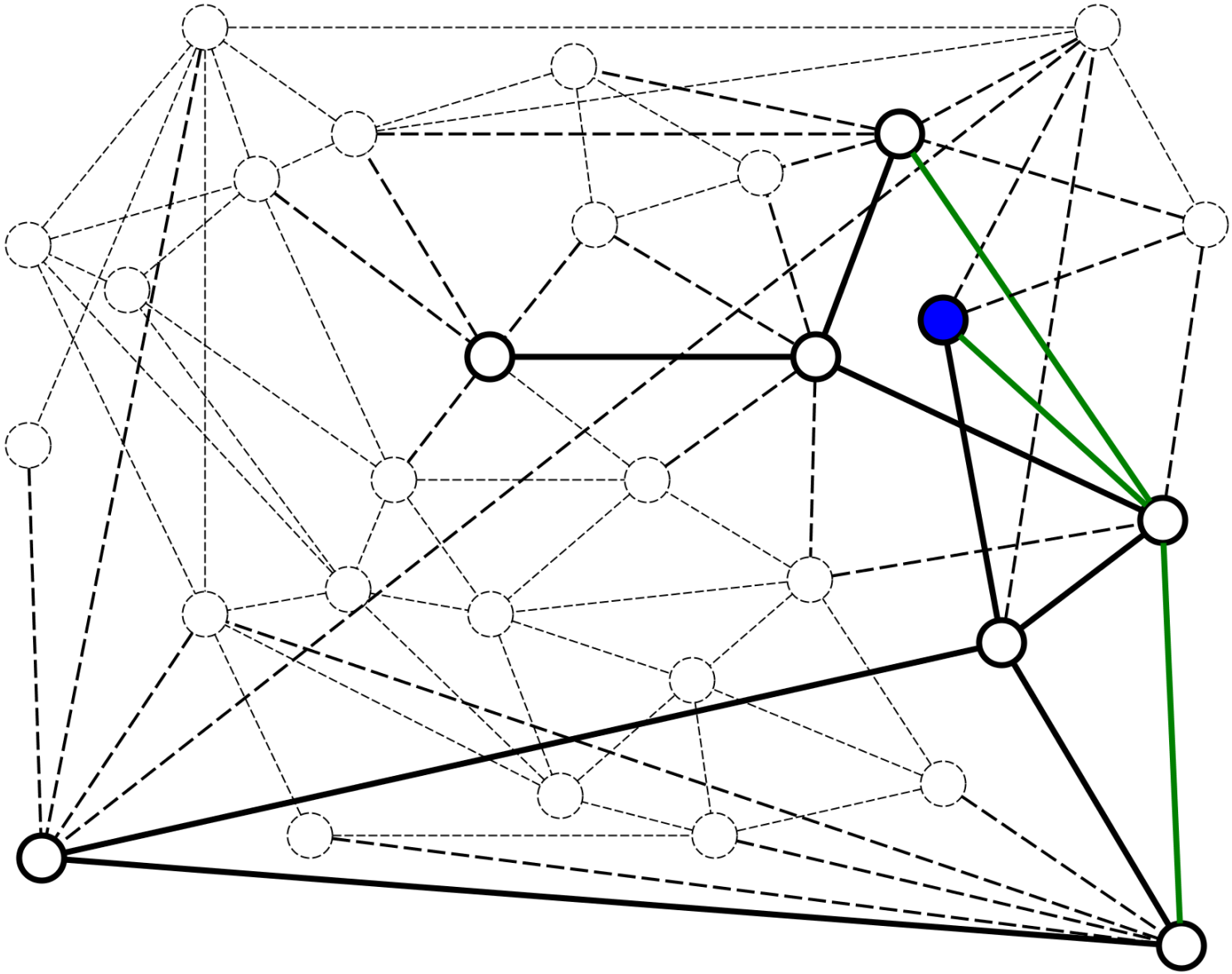


Методы сэмплинга. Исследование социальной сети

- Случайные блуждания:
  - Random Walk (RW):
    1. Случайный выбор инцидентного ребра
    2. Переход к соответствующей вершине
  - Random Jumps (RJ)
    - ★ Модификация: переход в случайную вершину с вероятностью p

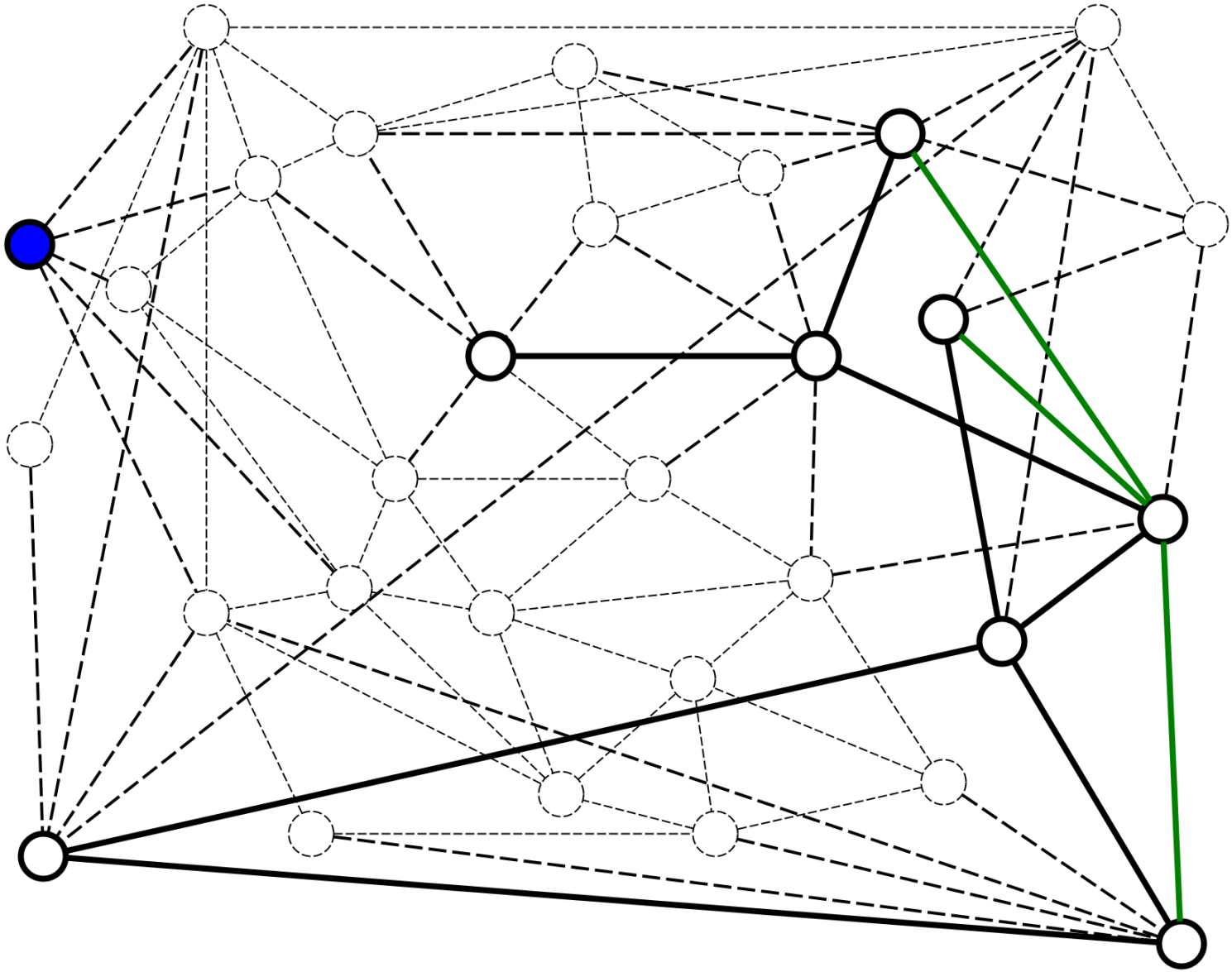
Методы сэмплинга. Исследование социальной сети

Случайные блуждания с Random Jumps (RJ). Пример



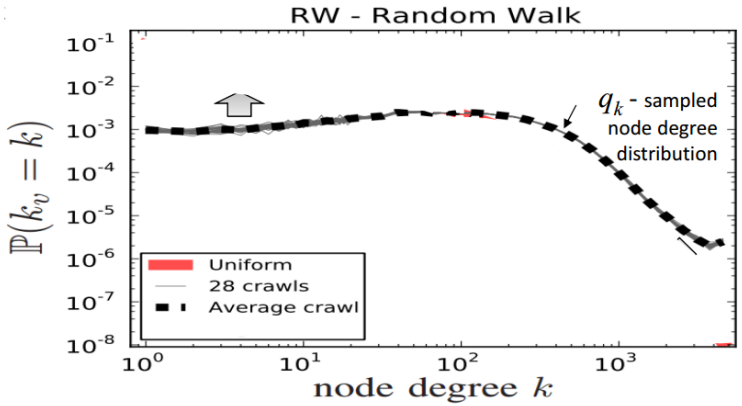
Методы сэмплинга. Исследование социальной сети

Случайные блуждания с Random Jumps (RJ). Пример



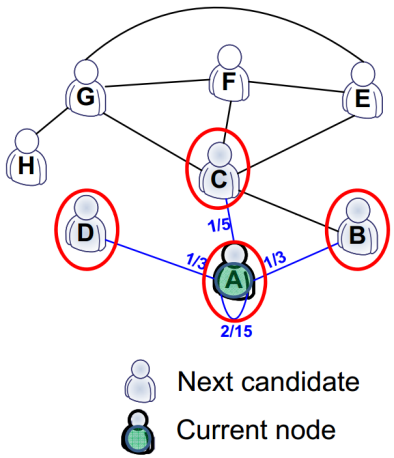
Методы сэмплинга. Исследование социальной сети

RW: Не получается степенное распределение



* http://mslab.csie.ntu.edu.tw/tut-pakdd13/samsum-pakdd13.pdf слайд 23

- Случайные блуждания:
  - Metropolis-Hastings Random Walks (MHRW) – решает данную проблему
    - ★ Вероятность перехода зависит от степеней вершин:



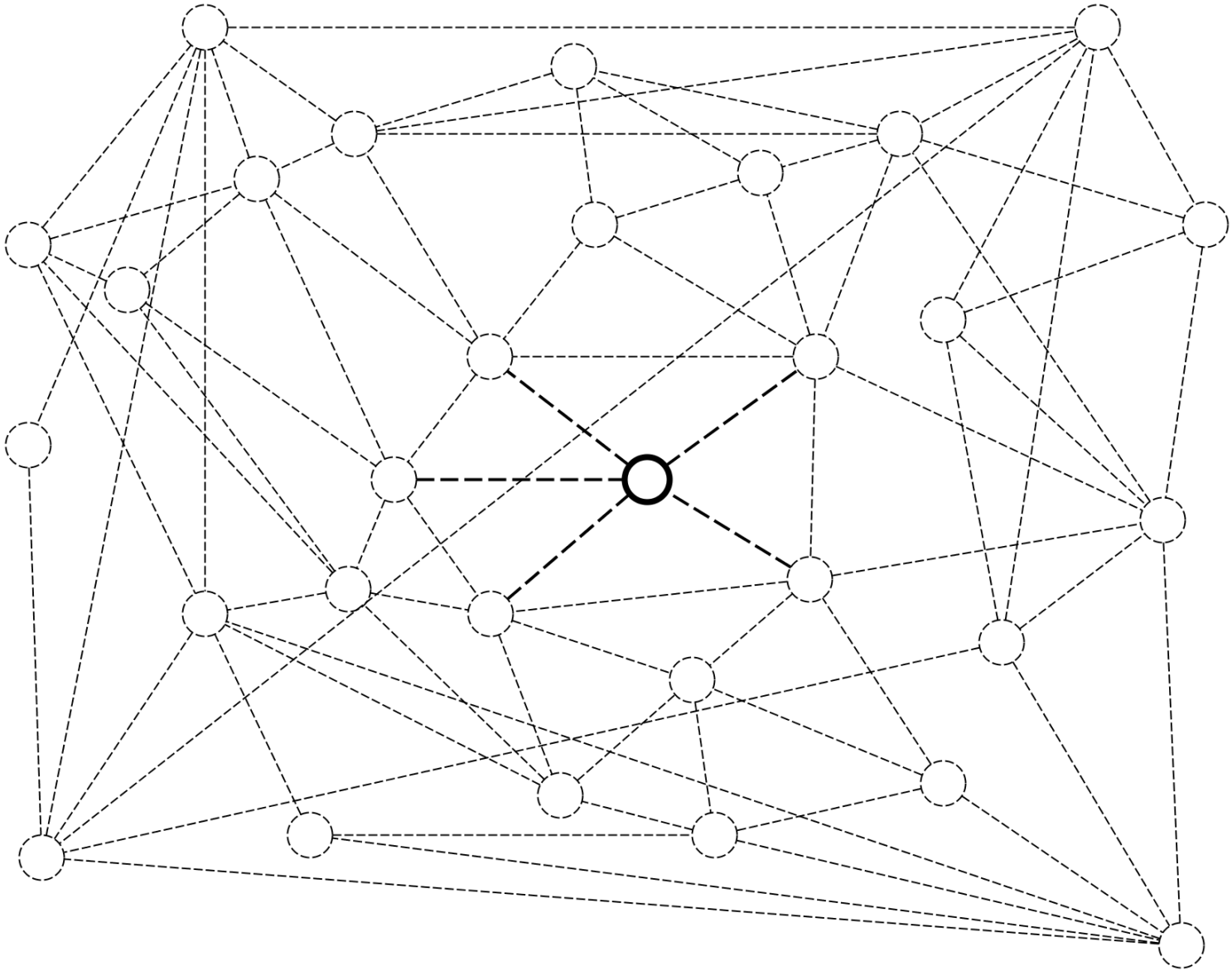
$$P_{v,w} = \begin{cases} \frac{1}{k_v} \times \min(1, \frac{k_v}{k_w}) & \text{если w и v – соседи} \\ 1 - \sum_{y \neq v} P_{v,y} & \text{если w=v} \\ 0 & \text{иначе} \end{cases}$$

Методы сэмплинга. Исследование социальной сети

- Поиск в ширину (BFS)
  - Обычно применяется, когда нужны локальные данные
  - Алгоритм:
    1. Добавить начальный узел в очередь
    2. Извлечь из начала очереди узел u и добавить в сэмпл
    3. В конец очереди добавляются все соседи узла u, которых ещё нет в сэмпле и которых еще нет в очереди
    4. Вернуться к п. 2

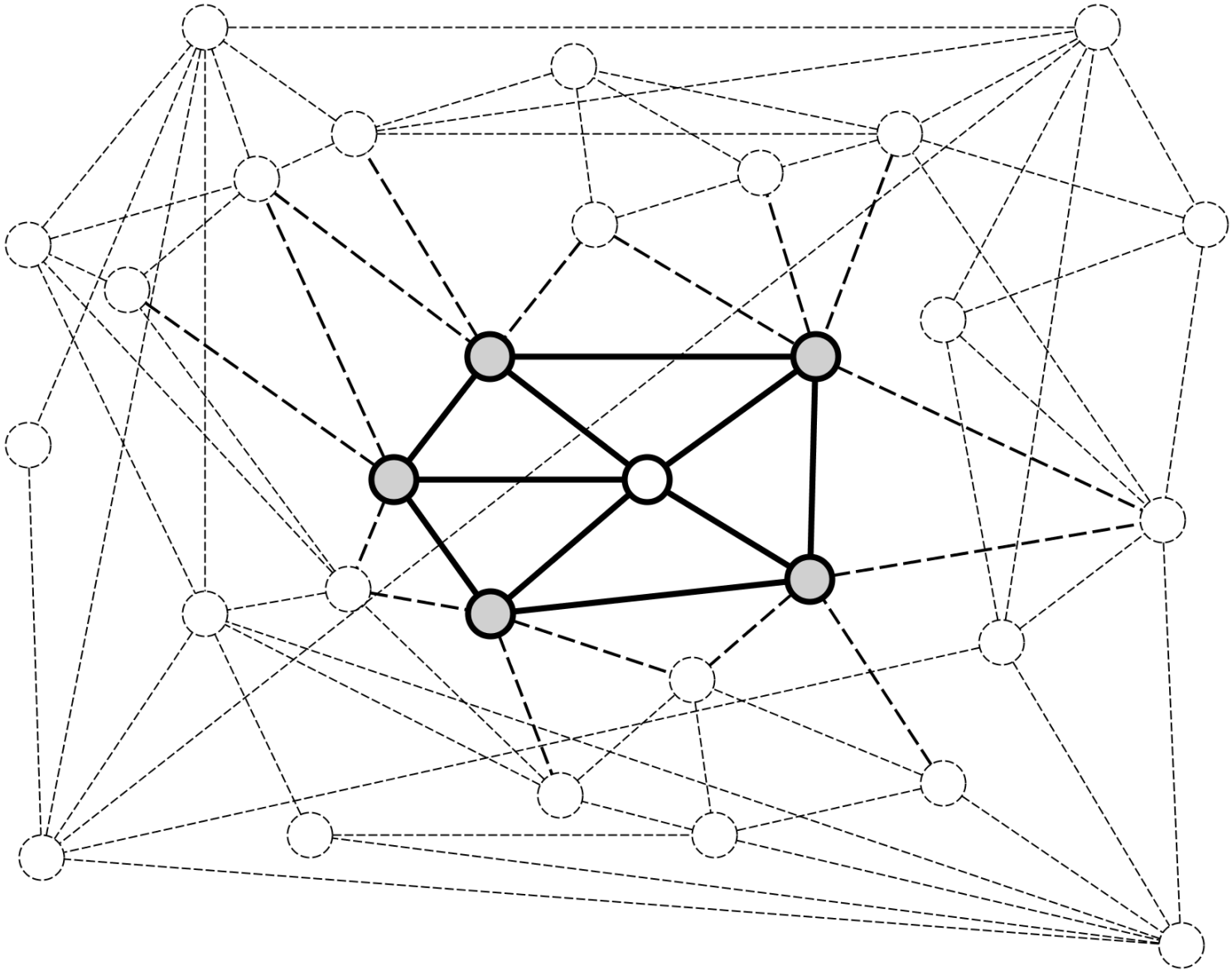
Методы сэмплинга. Исследование социальной сети

Поиск в ширину (BFS). Пример. Начальный пользователь



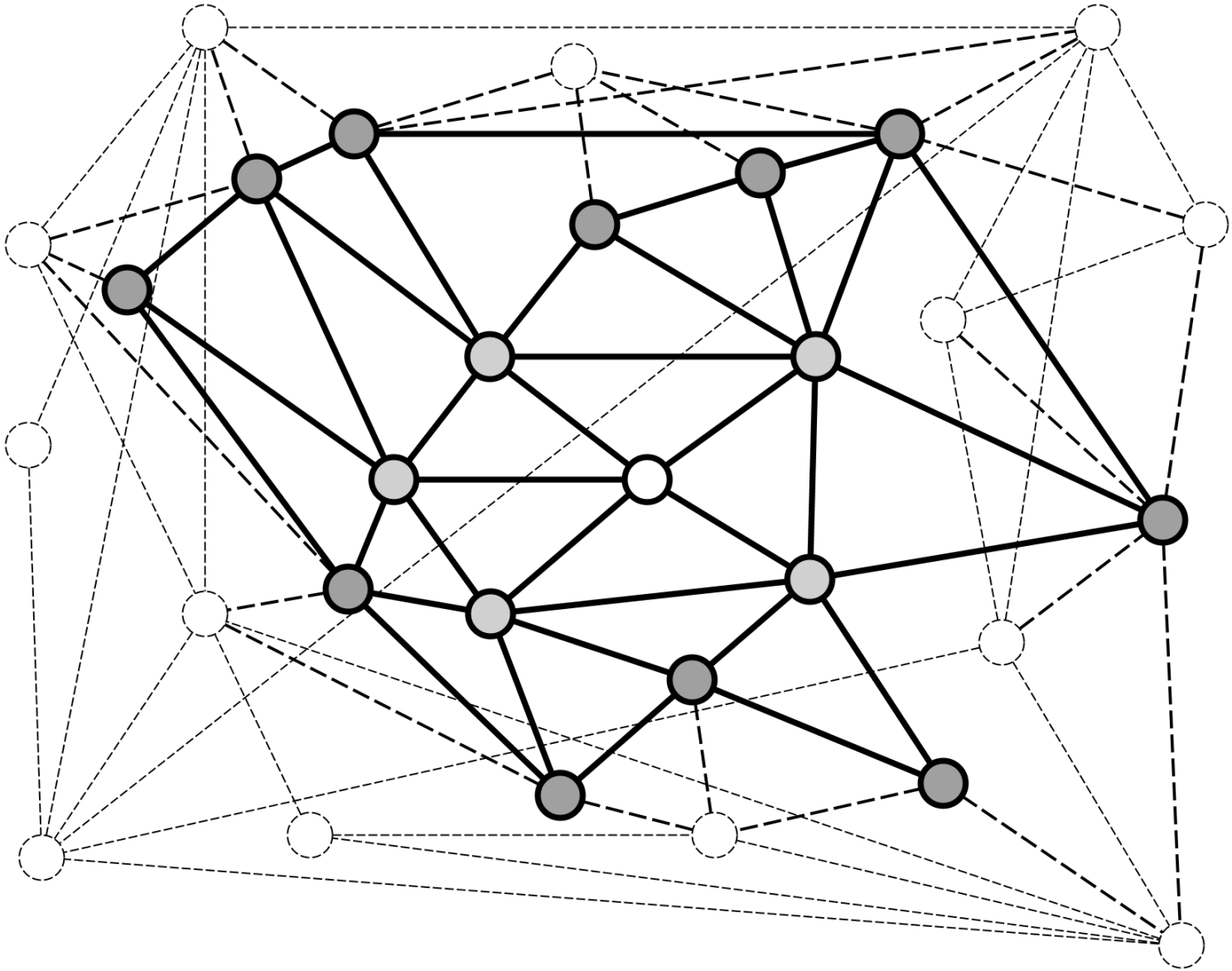
Методы сэмплинга. Исследование социальной сети

Поиск в ширину (BFS). Пример. 1-я окрестность



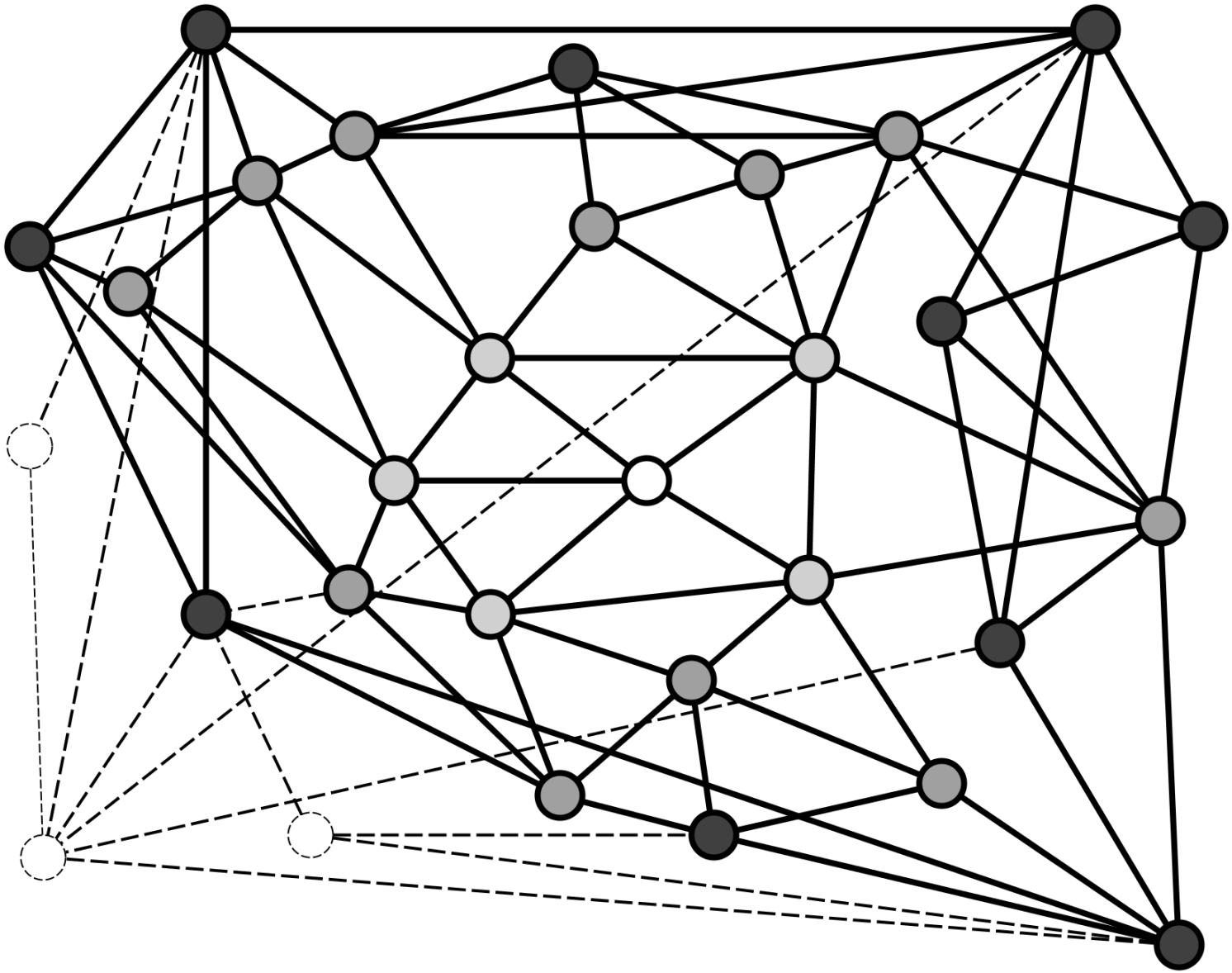
Методы сэмплинга. Исследование социальной сети

Поиск в ширину (BFS). Пример. 2-я окрестность



Методы сэмплинга. Исследование социальной сети

Поиск в ширину (BFS). Пример. 3-я окрестность

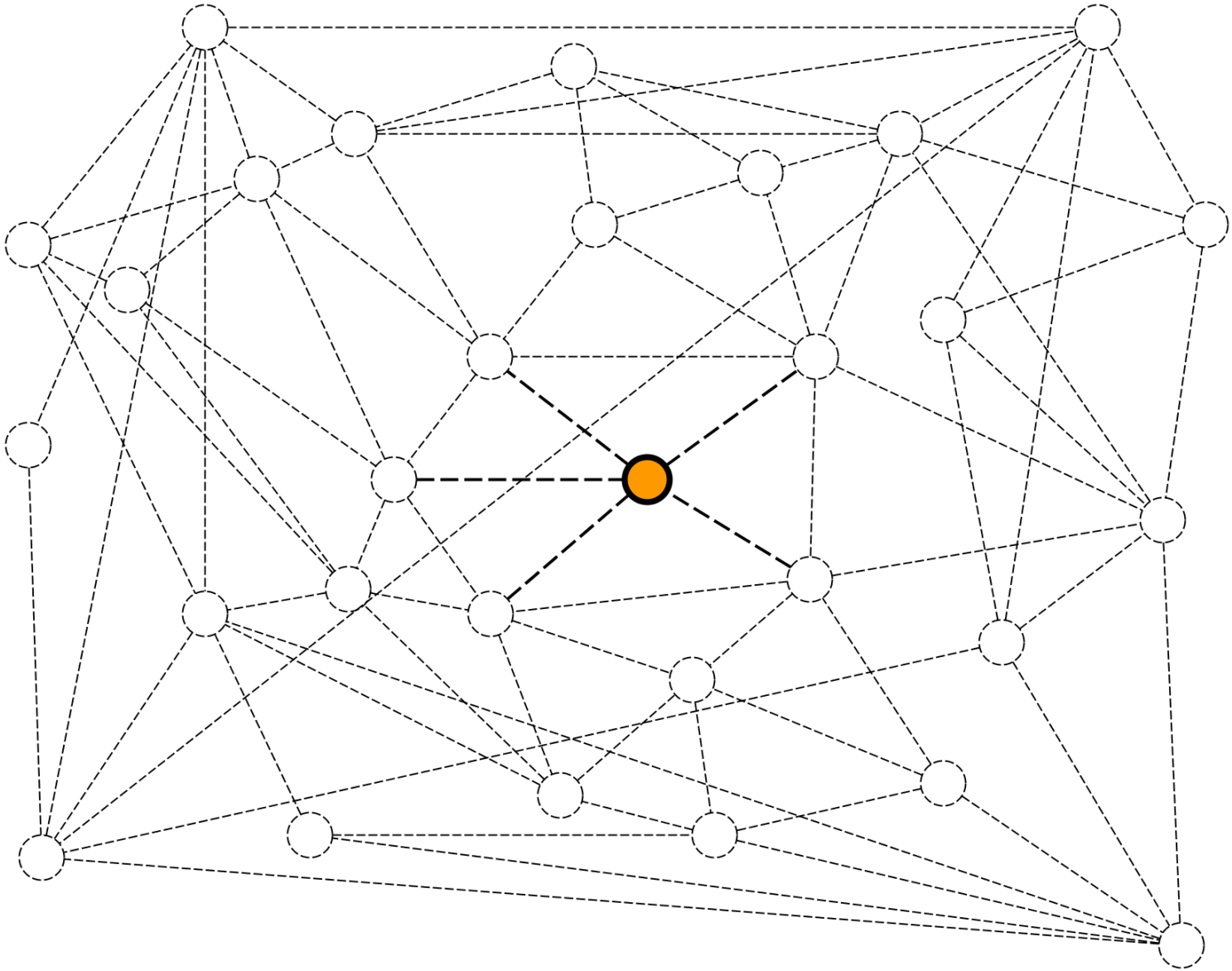


Методы сэмплинга. Исследование социальной сети

- Модель "Лесного пожара":
  1. Выбрать начальную вершину
  2. Если вершина еще "не сгорела":
     - ★ Пометить текущую вершину как "сгоревшую"
     - ★ Добавить вершину в сэмпл
     - ★ Выбрать соседей текущей вершины случайно, каждого независимо с вероятностью p
     - ★ Повторить п. 2 для выбранных соседей

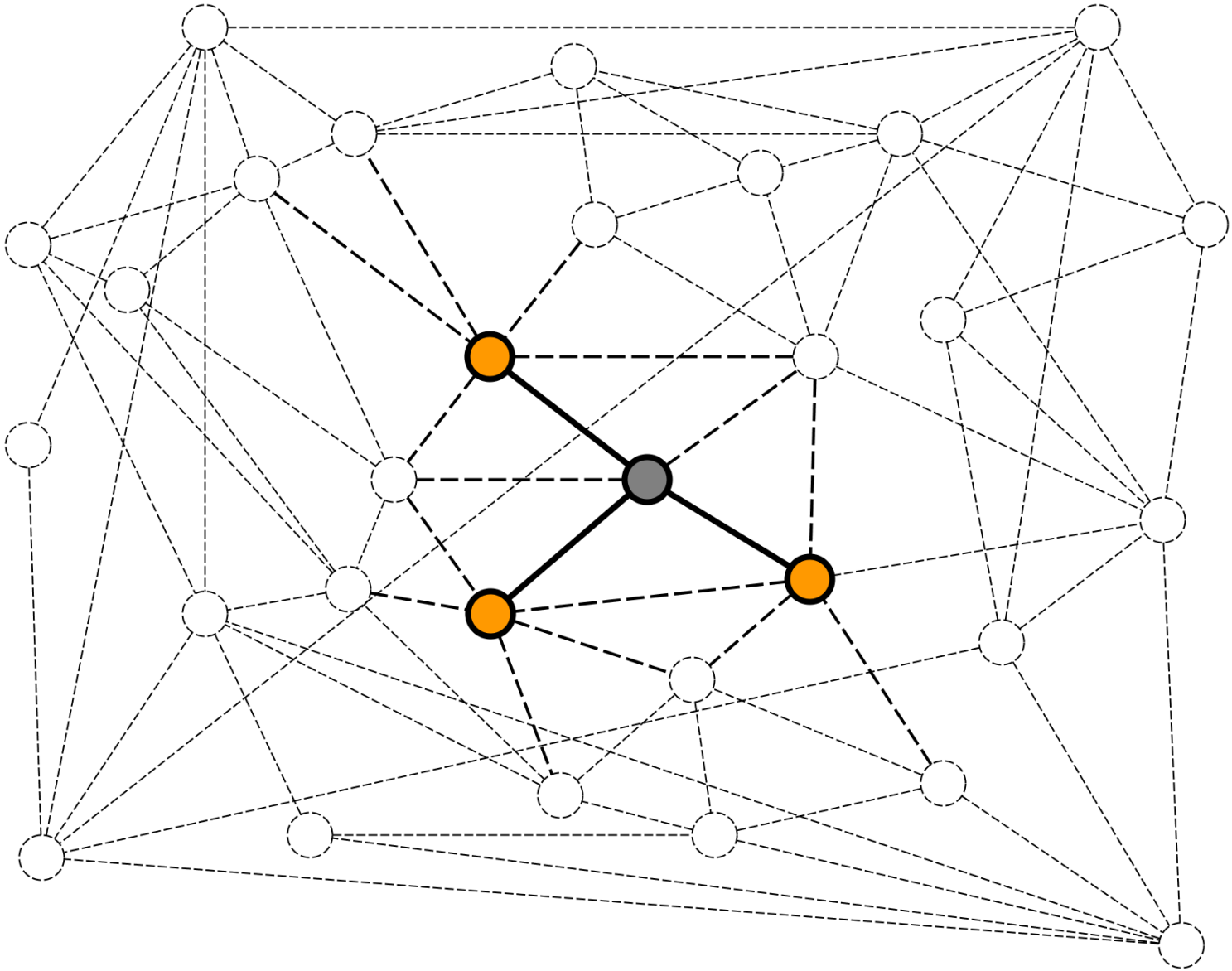
Методы сэмплинга. Исследование социальной сети

Лесной пожар (FF). Пример. Шаг 1



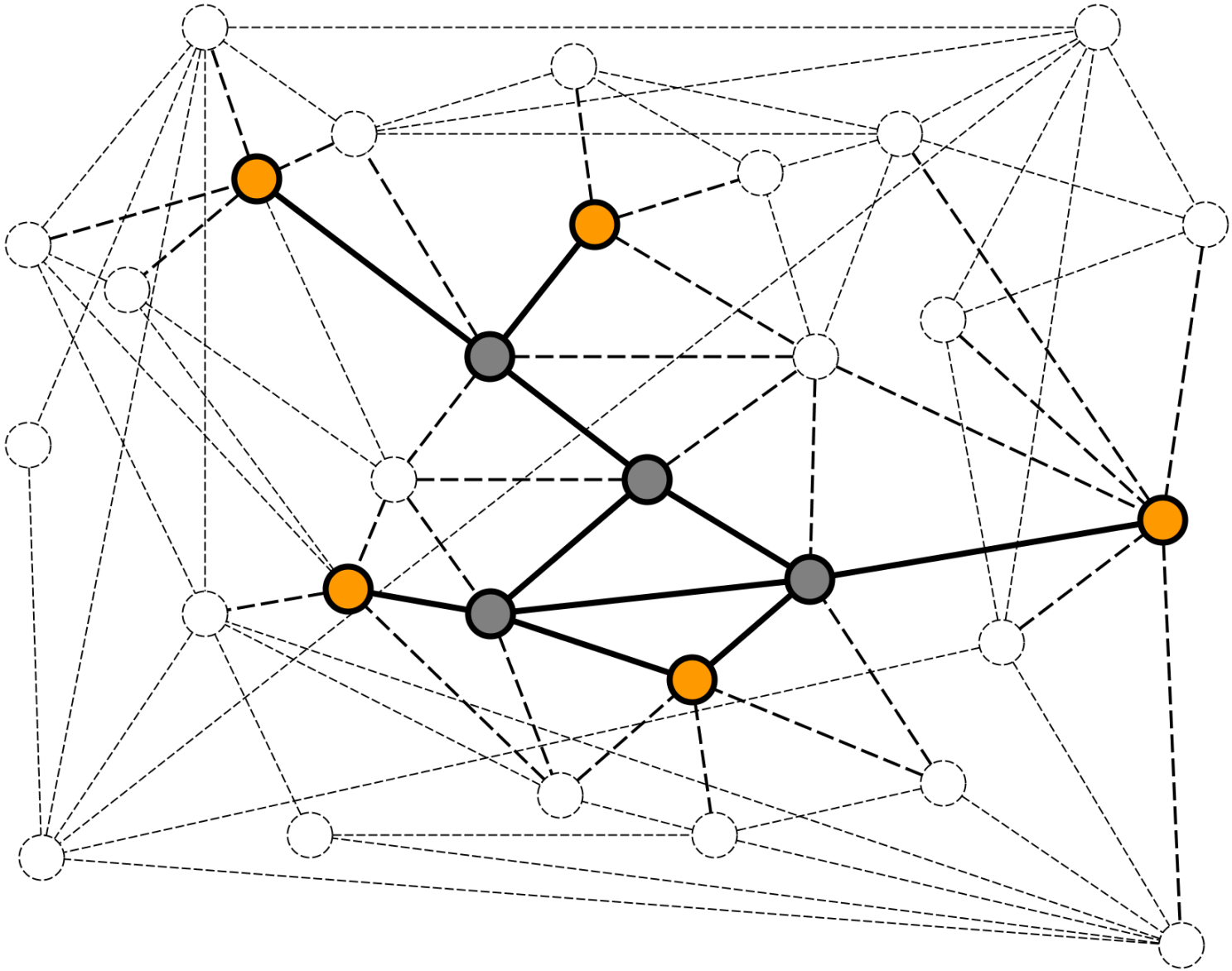
Методы сэмплинга. Исследование социальной сети

Лесной пожар (FF). Пример. Шаг 2



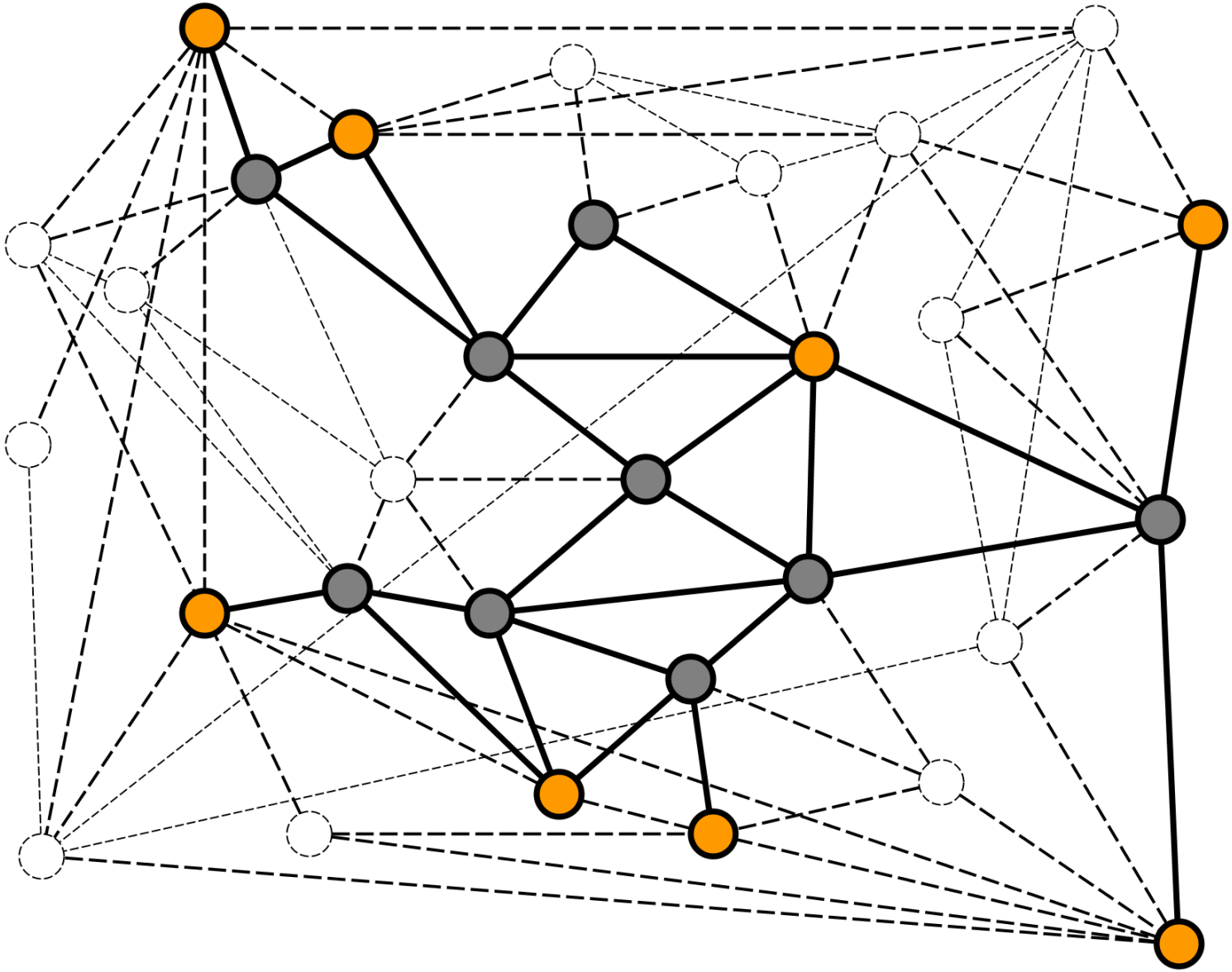
Методы сэмплинга. Исследование социальной сети

Лесной пожар (FF). Пример. Шаг 3



Методы сэмплинга. Исследование социальной сети

Лесной пожар (FF). Пример. Шаг 4



Методы сэмплинга. Исследование социальной сети

- Метод Mutual Friend Crawling
  - Идея: сэмпл - последовательность *сообществ*, а не пользователей
    - ★ Структура сообществ:
      Количество внутренних ребер (в пределах сообщества) больше, чем внешних (между разными сообществами)
  - Предполагается, что известны степени соседних вершин
  - Алгоритм:

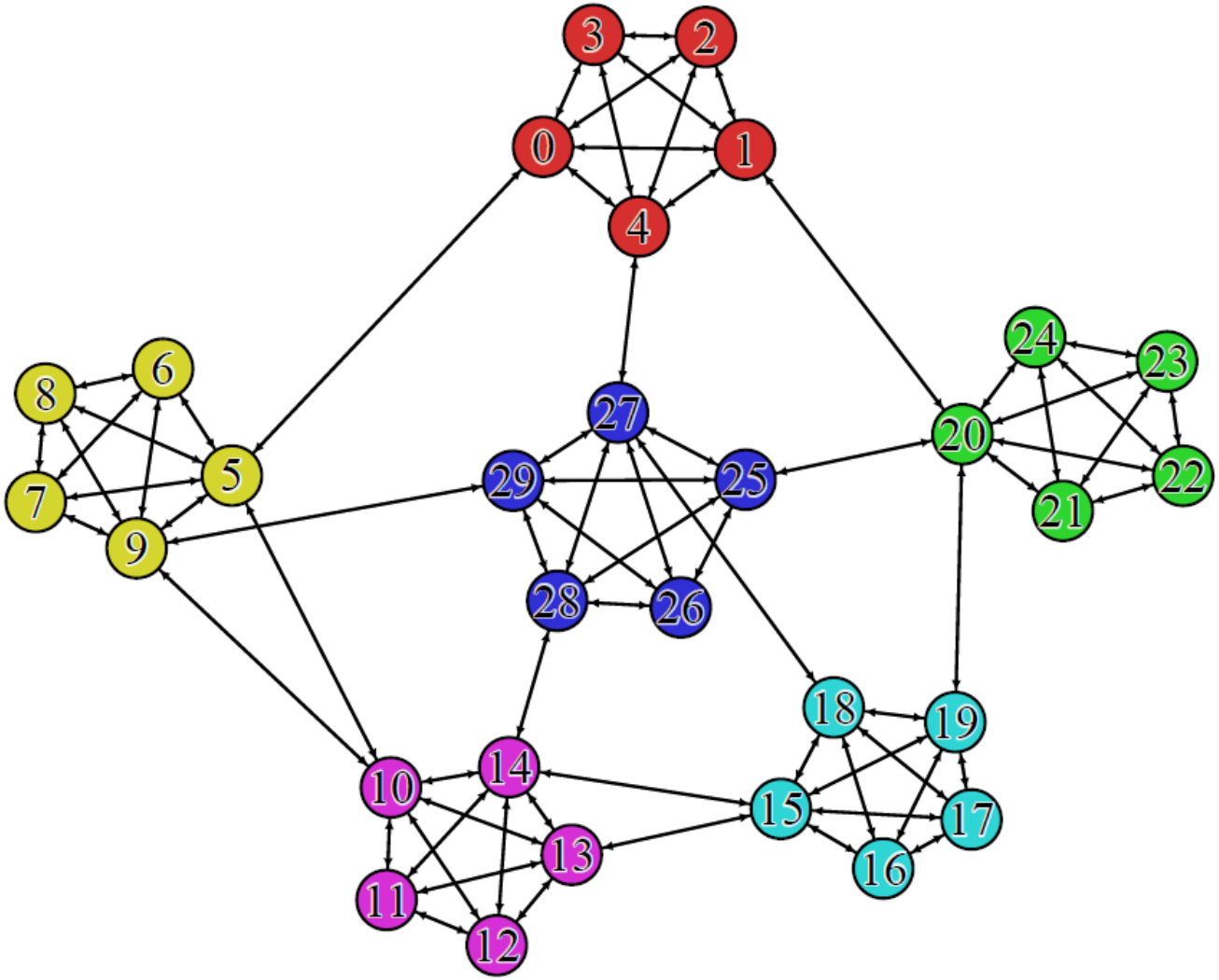
1. Выбрать первую вершину, добавить ее соседей в очередь
2. Для всех вершин из очереди вычисляется:

$$S_R = \frac{\text{количество ссылок на вершину}}{\text{степень вершины}}$$

3. Из очереди выбрать вершину с максимальным значением S_R
4. Добавить непросмотренных соседей выбранной вершины в очередь
5. Пока очередь не пуста повторять 2-5

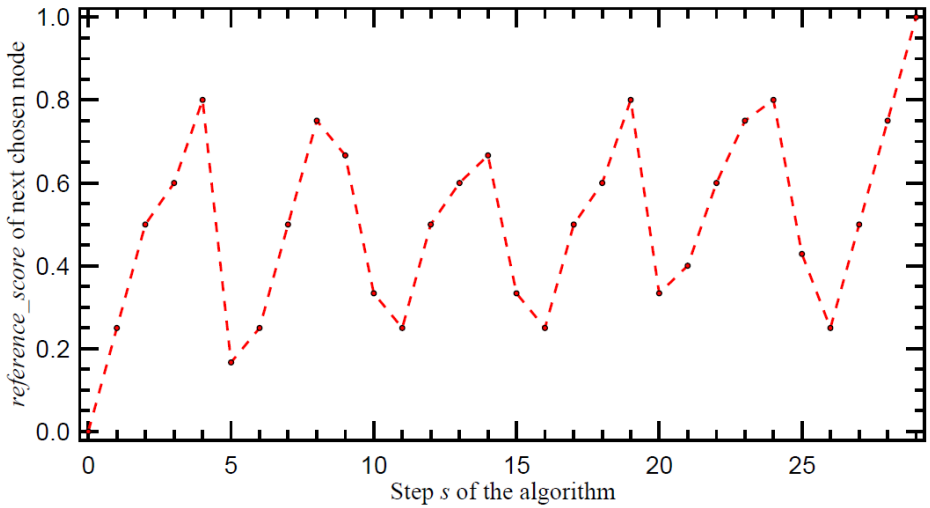
Методы сэмплинга. Исследование социальной сети

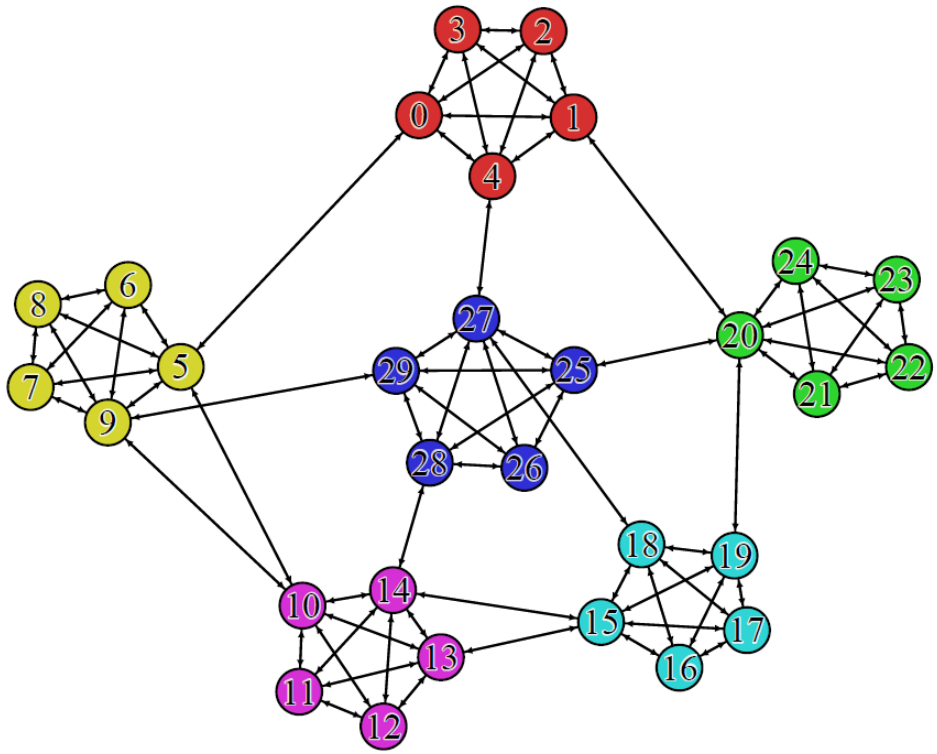
- Метод Mutual Friend Crawling



Методы сэмплинга. Исследование социальной сети

- Метод Mutual Friend Crawling
  - Данный метод позволяет автоматически в процессе получения выборки находить сообщества:





Методы сэмплинга. Резюме

- Выбор случайной вершины – надо знать id
  - Выбор из равномерного распределения (RN) – Не получается степенное распределение
  - Из распределения степени вершины (RDN) – Надо знать степень
  - Из распределения PageRank (RPN) – Надо знать PageRank
- Выбор случайного ребра – надо знать id
  - Random Edge (RE) – большой диаметр, несвязный граф
  - Random Node Edge (RNE)
  - Гибрид RE и RNE (HYB)
  - Induced Edge Sampling (IES)
- Исследование социальной сети
  - Random walks (RW)
    - ★ Модификация с рестартом (RWR)
    - ★ Модификация с "прыжками"(RJ)
    - ★ Метрополис-Гастингс (MHRW) – решает проблему распределения степени вершин
  - Поиск в ширину (BFS) – когда нужны локальные данные
  - Лесной пожар (FF)
  - Mutual Friend Crawling (сэмплинг по сообществам) – бонус: найденные сообщества

Оценка качества. D-статистика

Свойства (распределения):

1. In-degree
2. Out-degree
3. Размер WCC
4. Размер SCC
5. Hop-plot
6. Hop-plot на макс. WCC
7. Левый сингулярный вектор
8. Сингулярные значения
9. Коэффициент кластеризации

В работе (*) используется D-статистика Колмогорова-Смирнова для сравнения распределений:

$$D = \max_x \{|F'(x) - F(x)|\}$$

где F' и F – функции распределения для сэмпла и оригинального графа

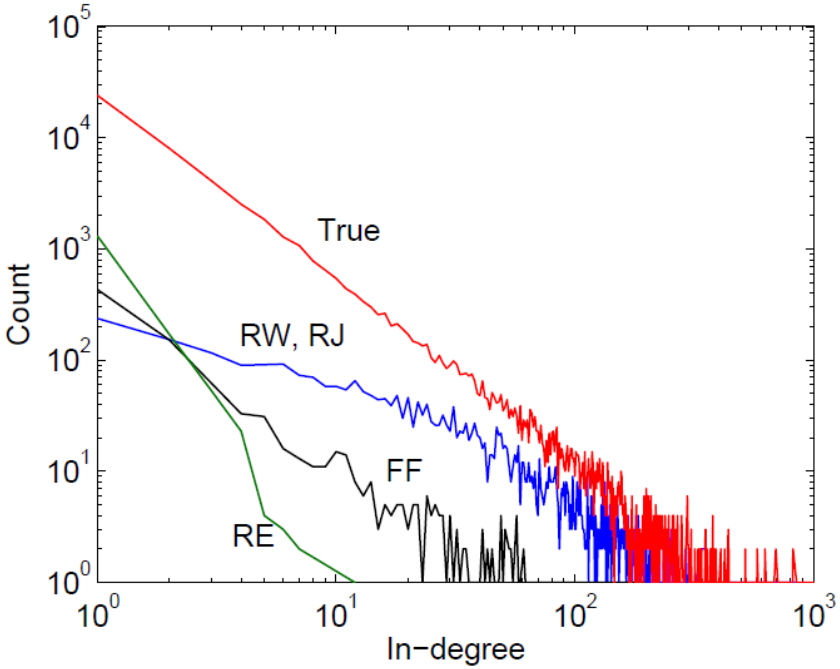
(*) Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

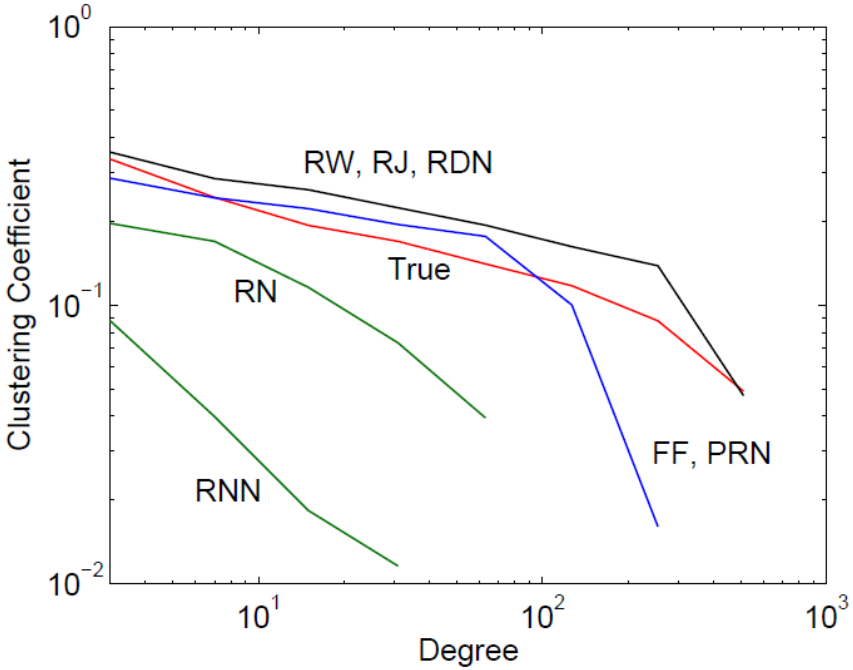
Оценка качества. Результаты

	in-deg	out-deg	wcc	scc	hops	sng-val	sng-vec	clust
RN	0.084	0.145	0.814	0.193	0.231	0.079	0.112	0.327
RPN	**0.062**	**0.097**	0.792	0.194	**0.200**	0.048	0.081	0.243
RDN	0.110	0.128	0.818	0.193	0.238	0.041	0.048	0.256
RE	0.216	0.305	**0.367**	0.206	0.509	0.169	0.192	0.525
RNE	0.277	0.404	0.390	0.224	0.702	0.255	0.273	0.709
HYB	0.273	0.394	0.386	0.224	0.683	0.240	0.251	0.670
RJ	0.132	0.151	0.771	0.215	0.264	0.076	0.143	**0.235**
RW	0.082	0.131	0.685	0.194	0.243	0.049	**0.033**	**0.243**
FF	0.082	0.105	0.664	0.194	**0.203**	**0.038**	0.092	**0.244**

(*) Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

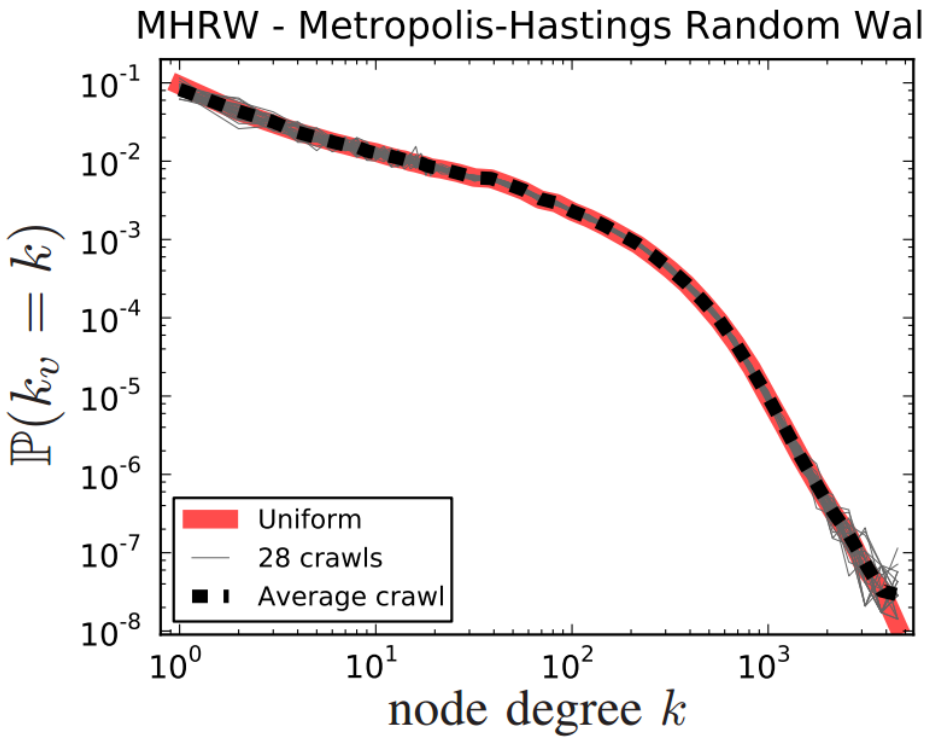
Оценка качества. Результаты

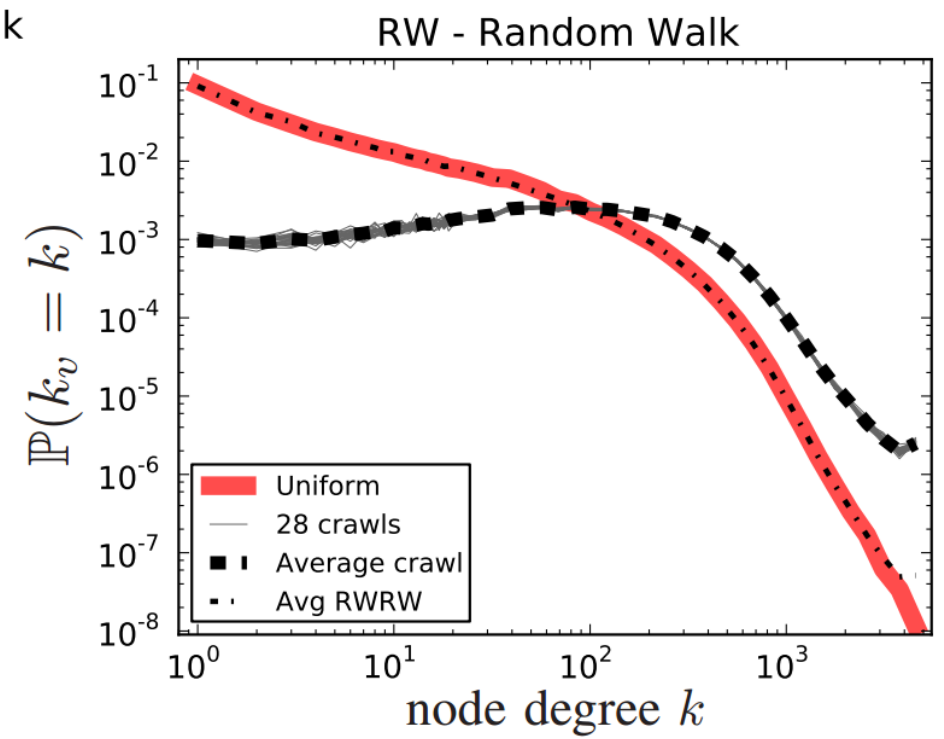




(*) Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

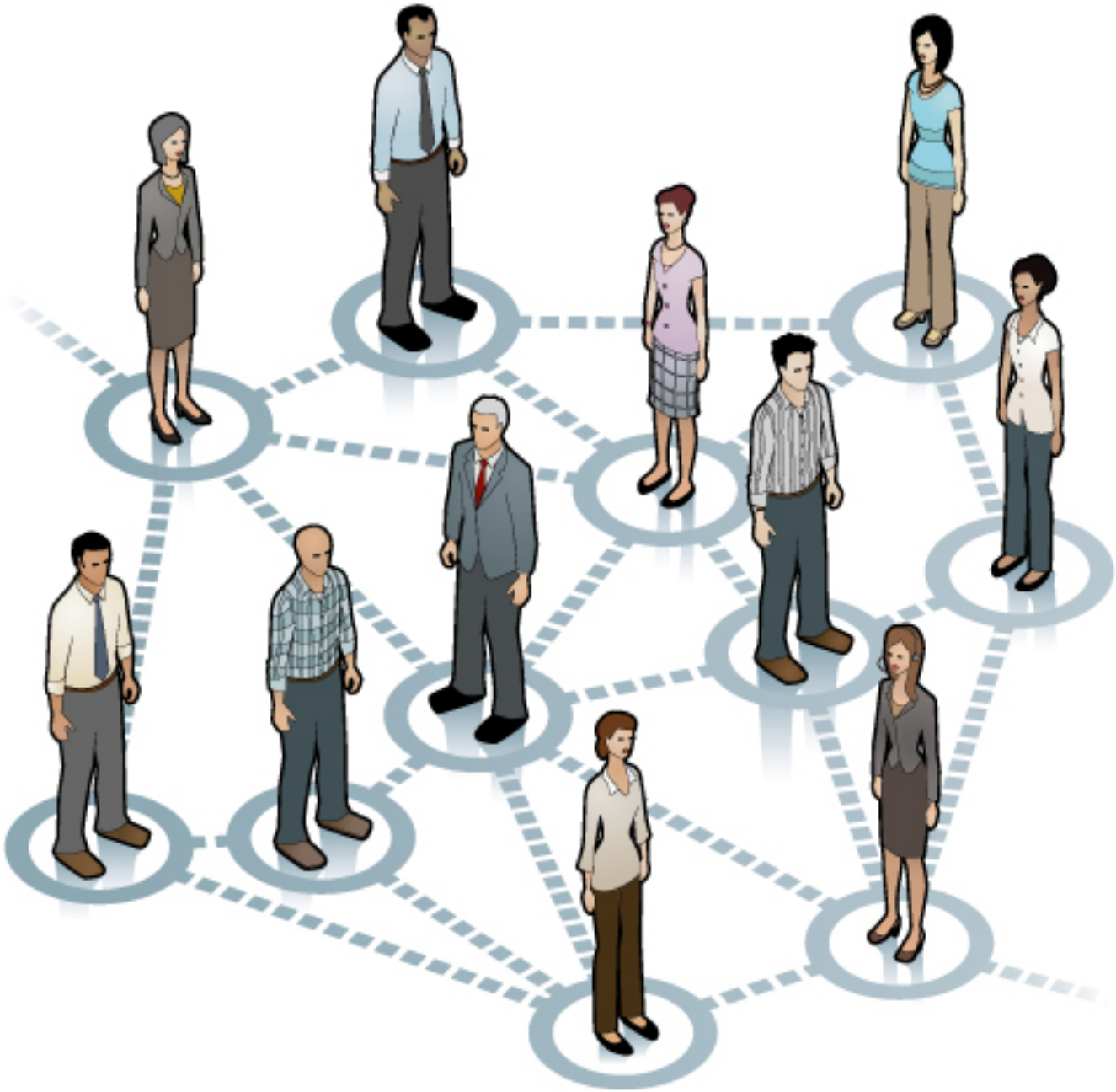
Оценка качества. Результаты (MHRW)





(*) Gjoka M. et al. Walking in facebook: A case study of unbiased sampling of osns //INFOCOM, 2010 Proceedings IEEE. – IEEE, 2010. – С. 1-9.

MODIS Crawler

- Социальные сети
  - Facebook (HTML)
  - Facebook (Graph API)
  - Twitter (API)
  - Hunch (HTML)
- Методы сэмплинга
  - BFS
  - FF
  - MHRW
- Борьба с банами и ограничениями
  - Автоматический выбор аккаунта для запроса
  - Использование одного аккаунта не чаще, чем раз в 5 секунд
  - Поддержка скачивания через прокси сервера
  - После привышения Rate limit аккаунт не используется в течение 15-минутного окна (Twitter API)
- MODIS Crawler как фреймворк
  - Можно реализовать свой метод сэмплинга

Спасибо за внимание