Яндекс

Каскады ретвитов – анализ и прогнозирование

Андрей Купавский

Исследователь, группа теории

Содержание доклада

1. Зачем предсказывать популярность контента?
2. Linear influence model
3. Машинное обучение
4. Контекст твита?

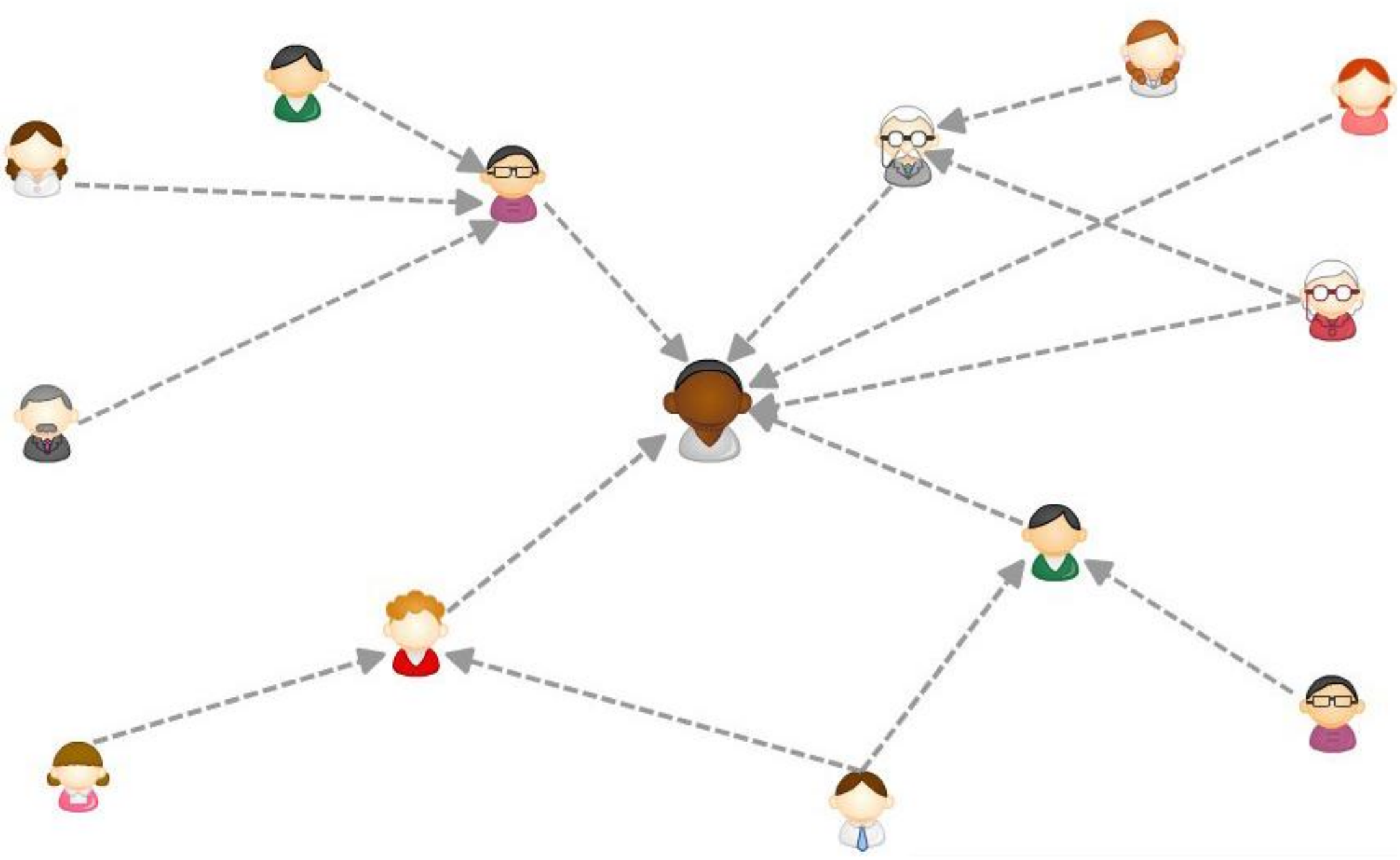
Зачем предсказывать популярность контента?

1. Реклама, продвижение брендов
2. Социальный журнализм
3. Улучшение качества и свежести выдачи

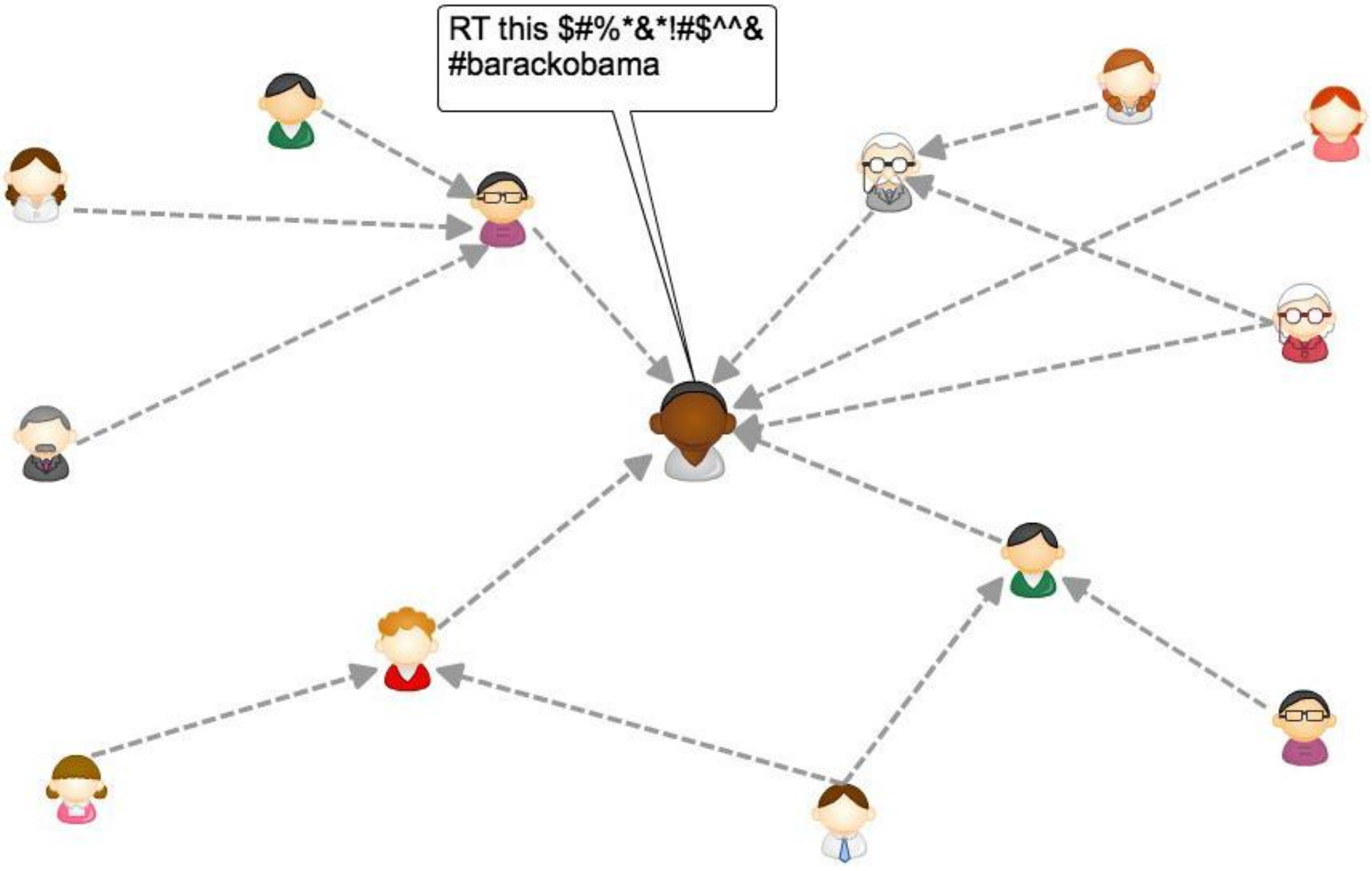
Предсказание популярности в Твиттере

- Ретвиты
- Показы
- Клики

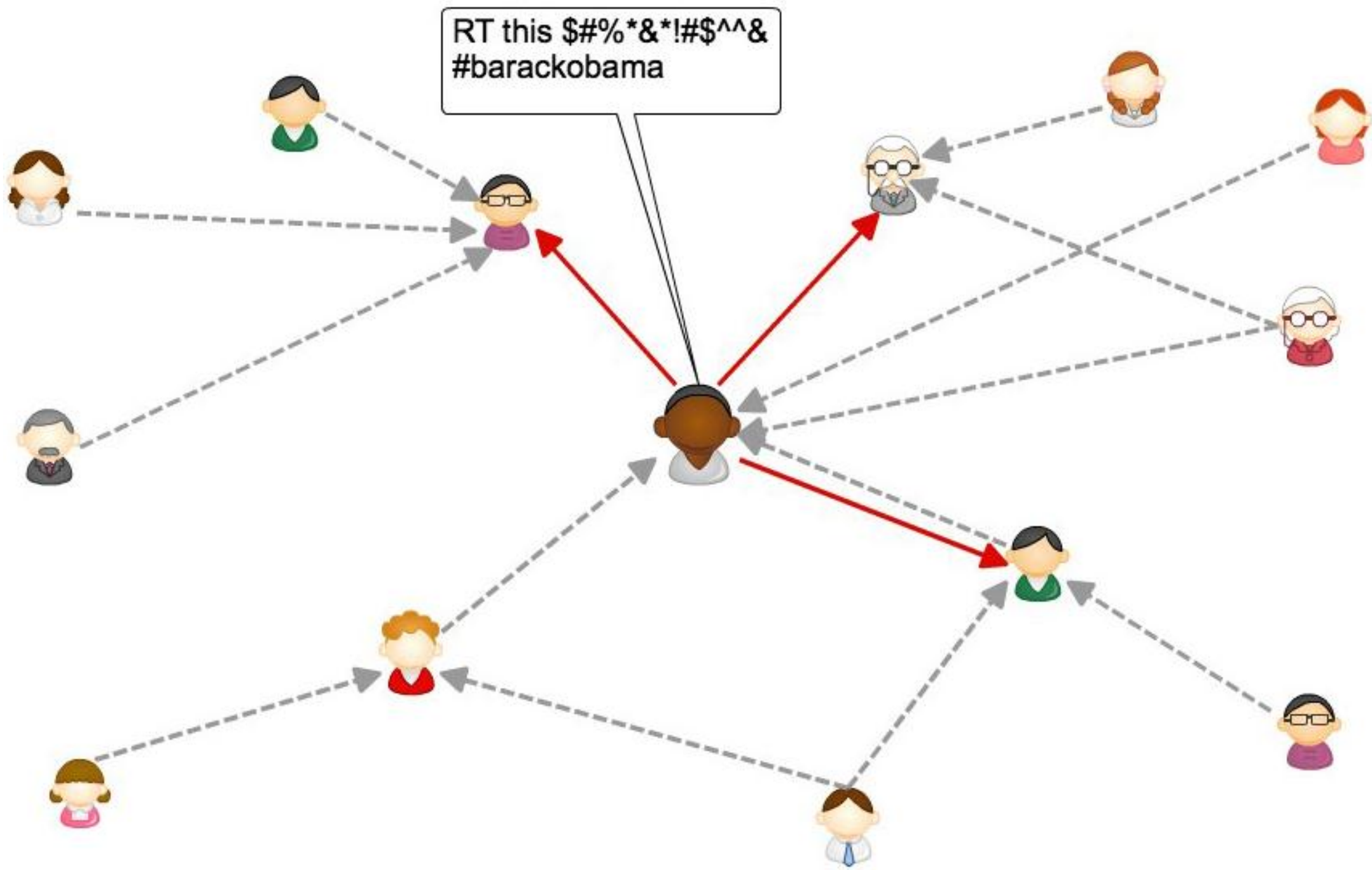
Ретвит-каскады



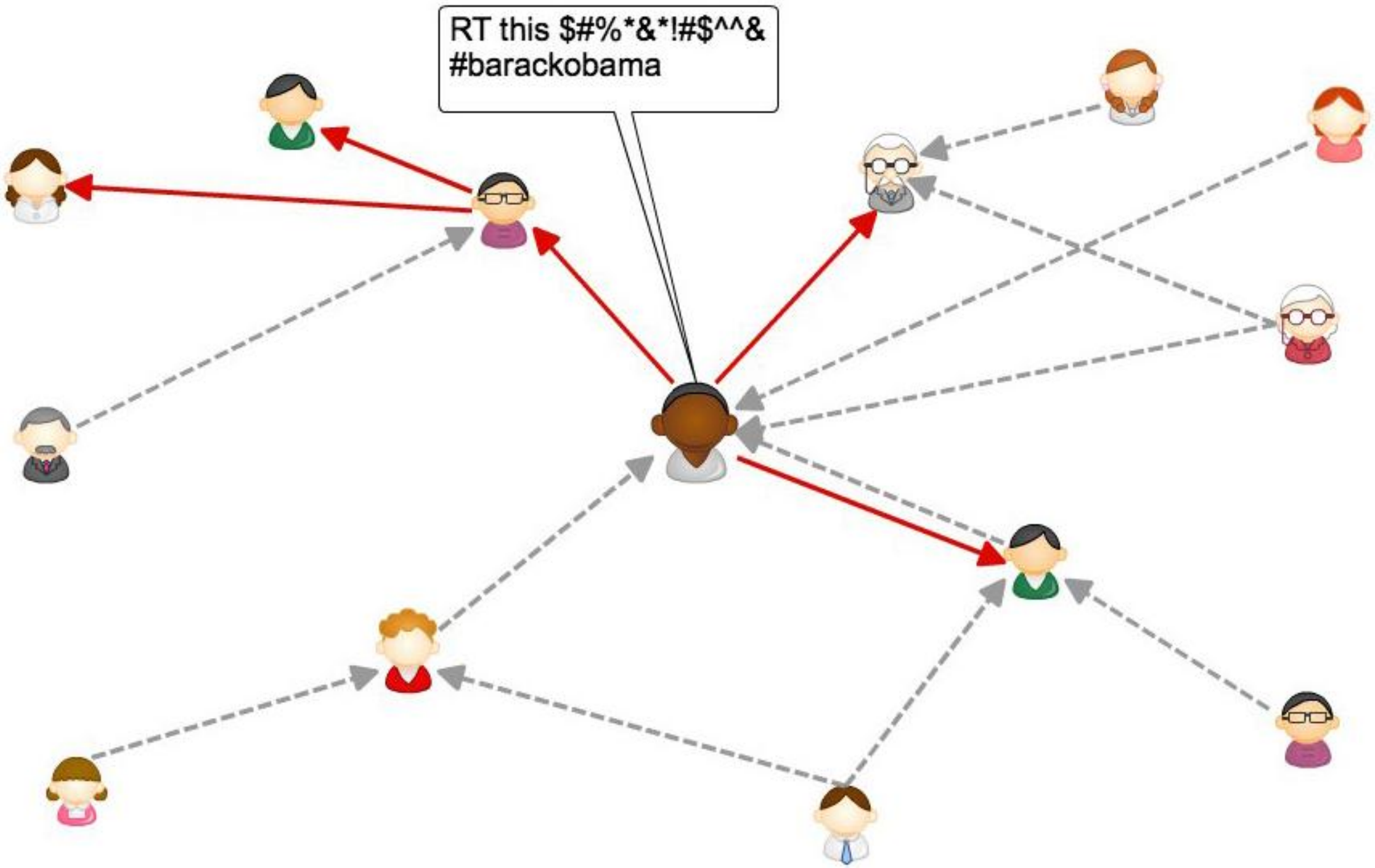
Ретвит-каскады



Ретвит-каскады



Ретвит-каскады



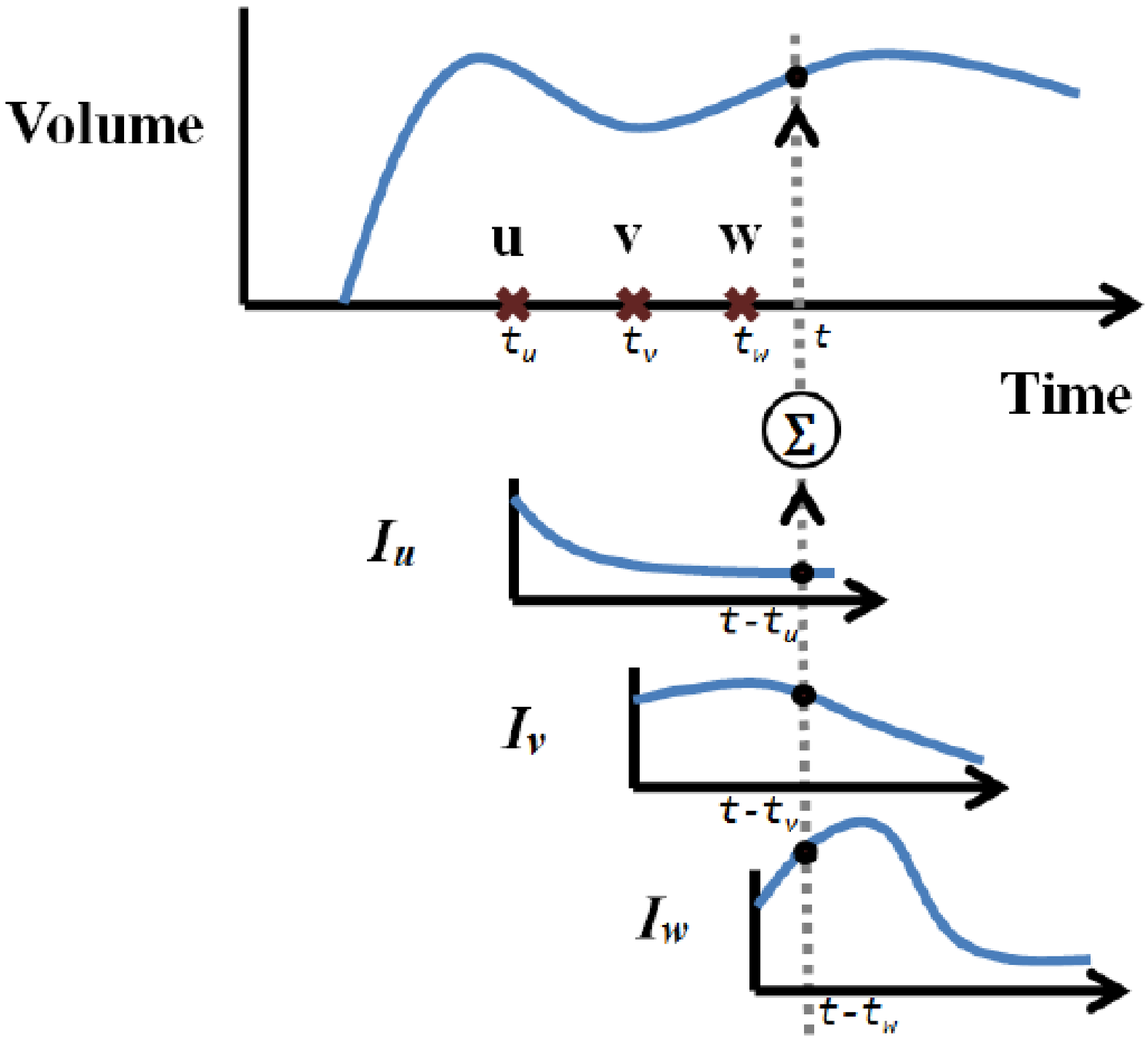
Ретвит-каскады

- ретвиты получает **5-6 %** всех сообщений
- Популярность распределена в соответствии со степенным законом: число твитов, получивших k ретвитов, пропорциональна величине $1/k^c$ для некоторой константы c
- за первый час происходит **90 %** всех ретвитов

Модели распространения информации

1. Linear influence model
2. Машинное обучение
3. Multiple interaction model

Linear influence model



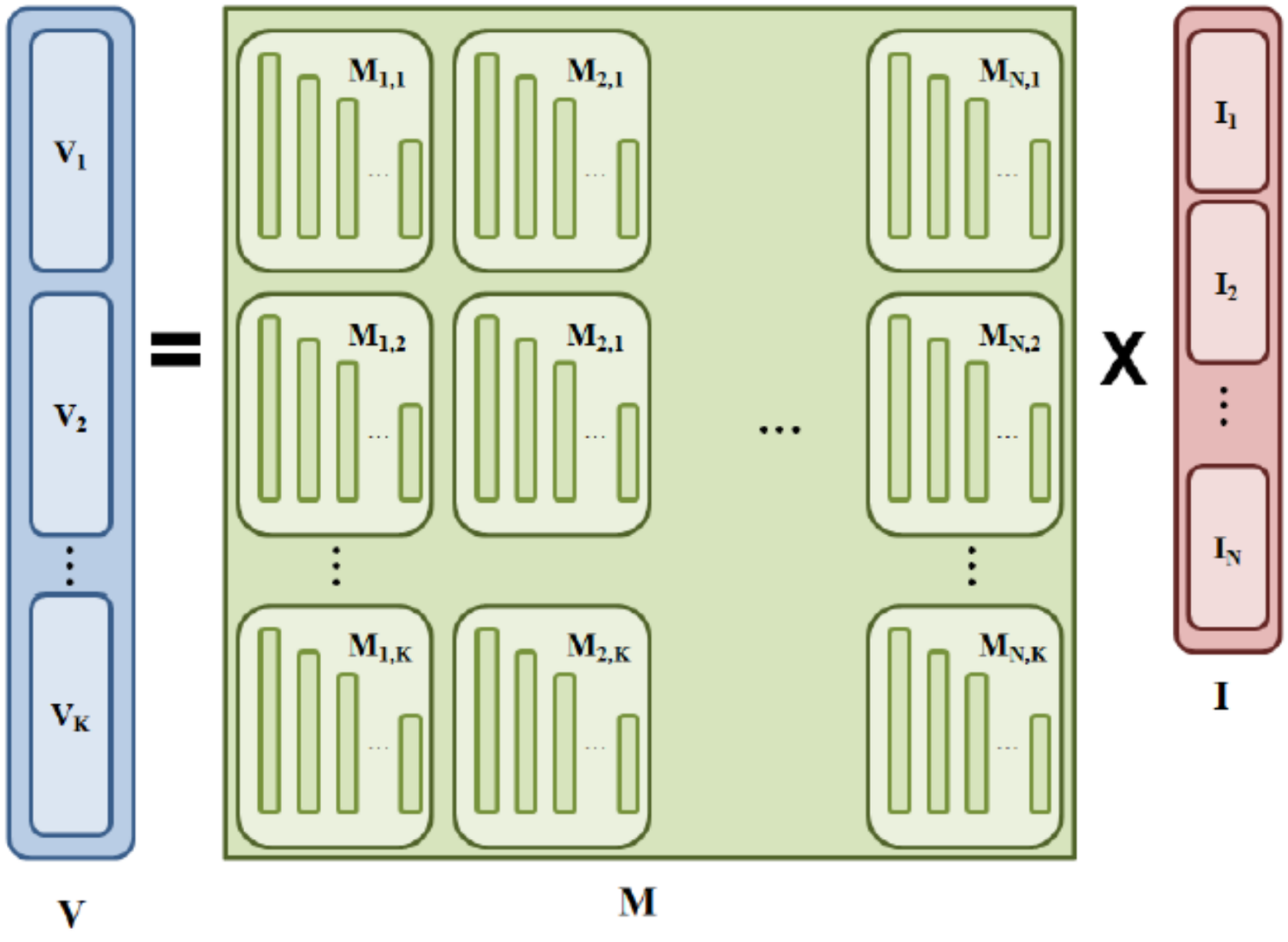
Linear influence model

- Зная распространение на момент i, предсказываем объем на момент i+1
- Фиксированные источники с функцией влиятельности, зависящей от времени
- Объем – сумма влиятельностей зараженных до этого источников:

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u)$$

Linear influence model

Для того, чтобы вычислить функцию влияния, нужно приближенно решить систему линейных уравнений, описывающих предыдущие каскады.



Анализ

+ :

- влияние источников явно вычисляется
- влияние зависит от времени
- не требуется структура сети

– :

- фиксированные источники
- вычислительная сложность
- не учитывается структура сети
- предсказание только на следующий момент

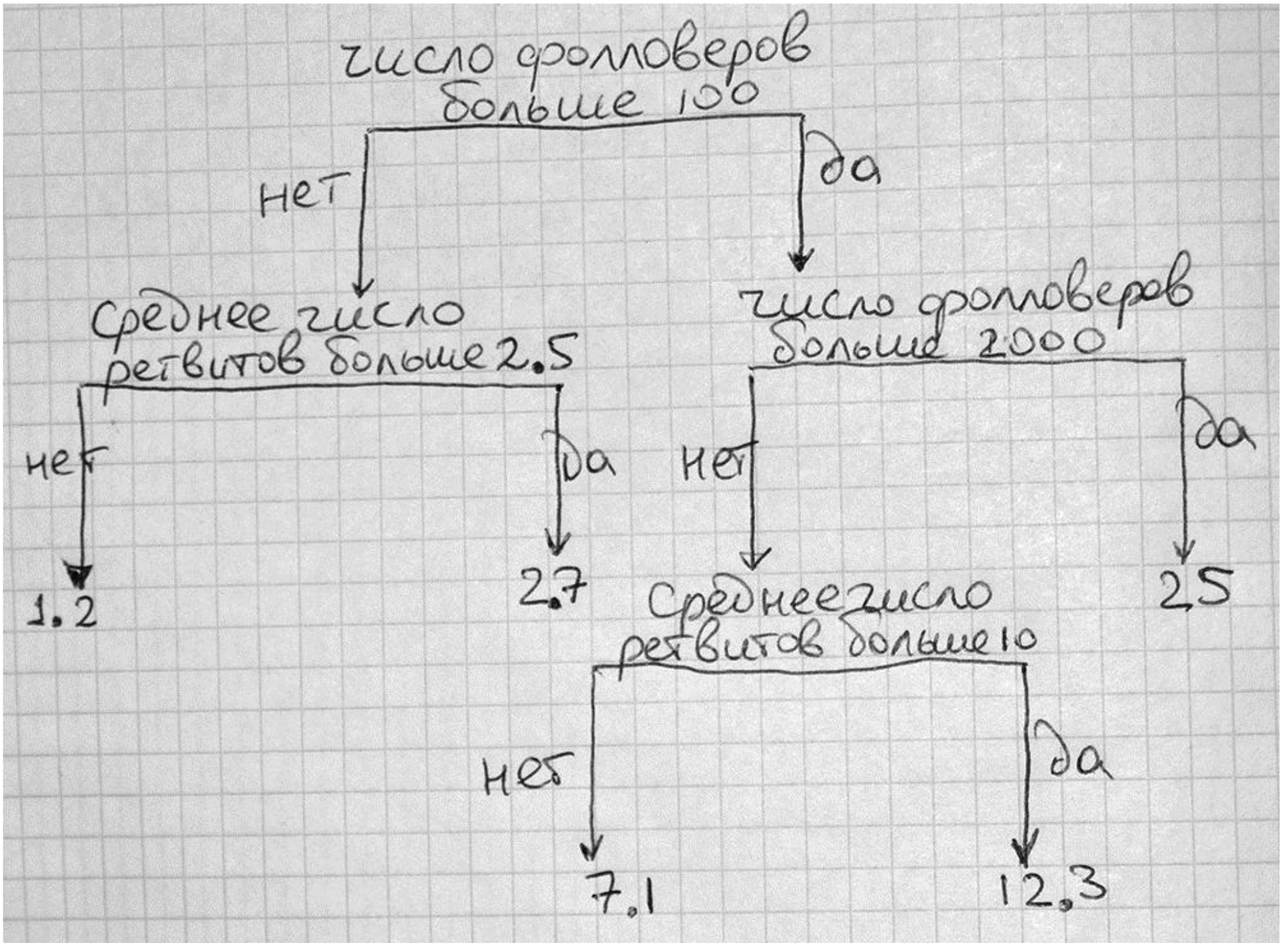
Анализ

- Подходит для моделирования распространения информации в блогах и СМИ
- Для задачи предсказания распространения в Твиттере плохо, что: мало источников, краткосрочное предсказание

Машинное обучение

- Есть целевая функция
- Есть обучающая выборка, на ней вычисляются факторы
- Алгоритм оптимальным образом разбивает твиты на классы. В каждом классе предсказанная популярность постоянна

Машинное обучение



Анализ

+ :

- малая вычислительная сложность
- дают хороший результат
- можно улучшать за счет новых факторов

– :

- отсутствие физической модели
- непонятно, как именно распространяется сообщение

Машинное обучение

1. Факторы
2. Экспериментальные результаты
3. Важность факторов

Социальные факторы

- Число читателей
- Число читаемых
- Среднее число ретвитов
- Дата создания аккаунта
- …

Контентные факторы

- Длина сообщения
- Наличие хештегов
- Наличие ссылок
- Является ли сообщение ответом
- Настроение
- …

Начальное распространение

- Число ретвитов за начальный период
- Авторитетность пользователей, сделавших ретвит
- …

Экспериментальные результаты

- Предсказание точнее на короткие промежутки
- Предсказание на 20-30% точнее с данными за первые 30 секунд
- Текстовые факторы почти не играют роли
- Предсказанное число ретвитов отличается от реального в 2-3 раза.

Важность факторов

- Популярность твита в первую очередь зависит от авторитетности пользователя
- Начальное распространение определяет, насколько удачен этот твит для пользователя

Multiple interaction model

- Модель позволяет предсказывать для каждого пользователя, опубликует он или нет тот или иной твит.
- При предсказании учитывается, какие твиты пользователь видел перед этим

Multiple interaction model

- Для каждого пользователя и твита вычисляется вероятность того, что он этот твит опубликует
- Если вероятность выше фиксированного порога, то считается, что пользователь твит опубликовал
- Влияние последних K твитов раскладывается в произведение влияний каждого с учетом его позиции:

$$P\left(X|\left\{Y_k\right\}_{k=1}^{K}\right) = \frac{1}{P(X)^{K-1}}\prod_{k=1}^{K} P(X|Y_k).$$

Multiple interaction model

- Все твиты классифицируются на небольшое количество скрытых тем
- По факту мы учитываем именно то, как взаимодействуют между собой темы.
- В итоге вероятность публикации твита складывается из заразности твита, восприимчивости пользователя и суммы влияний тем предыдущих твитов:

$$P_n(X = u_j | Y_k = u_i) = P(X = u_j) + \gamma_n + \sum_t \sum_s \mathbf{M}_{i,t} \cdot \Delta_{t,s}^{(k)} \cdot \mathbf{M}_{j,s}$$

Анализ

- Точность анализировалась на основе precision-recall.
- Учет контекста дает повышение точности в несколько раз
- Слагаемые в вероятности публикации твита, отвечающие за контекст, дают 70% вероятности

Анализ

- Контекст твита играет значительную роль в предсказании распространения каскада
- Модель вычислительно очень затратная, поэтому имеет скорее теоретический характер

Дальнейшие исследования

- Использовать информацию о контексте твита в машинном обучении
- Применение в различных сервисах Яндекса

Литература

- A. Kupavskii et. al., “Prediction of Retweet Cascade Size over Time”, CIKM'12
- A. Kupavskii et. al., “Predicting the Audience Size of a Tweet”, ICWSM'13
- H. Kwak et. al., “What is Twitter, a Social Network or a News Media?”, WWW'10
- S. A. Myers, J. Leskovec, “Clash of the Contagions: Cooperation and Competition in Information Diffusion”, ICDM’12
- J. Yang, J. Leskovec, “Modeling Information Diffusion in Implicit Networks”, IEEE'10