МОДЕЛИ СОЦИАЛЬНЫХ СЕТЕЙ

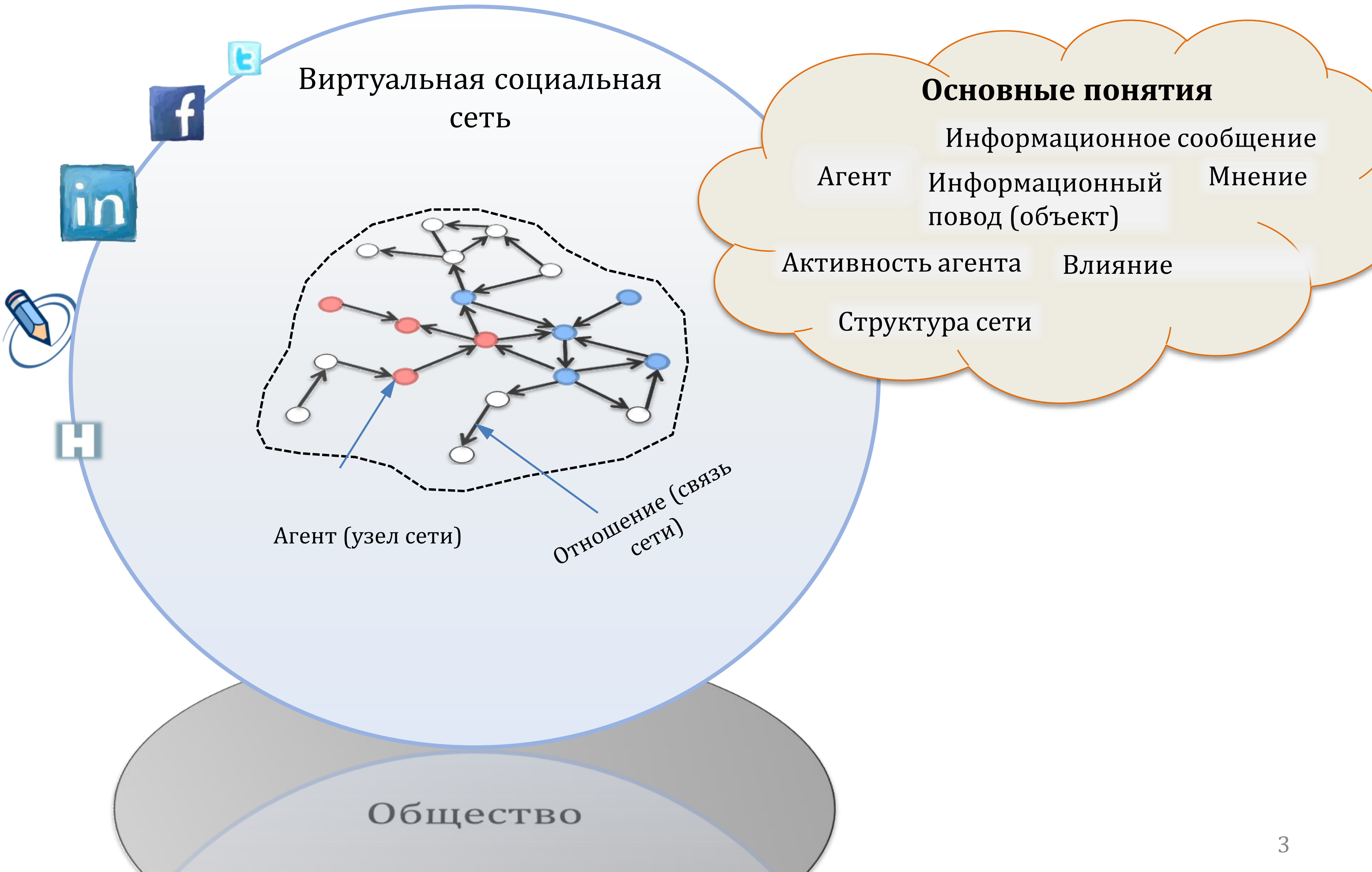
Губанов Дмитрий Алексеевич

старший научный сотрудник ИПУ РАН

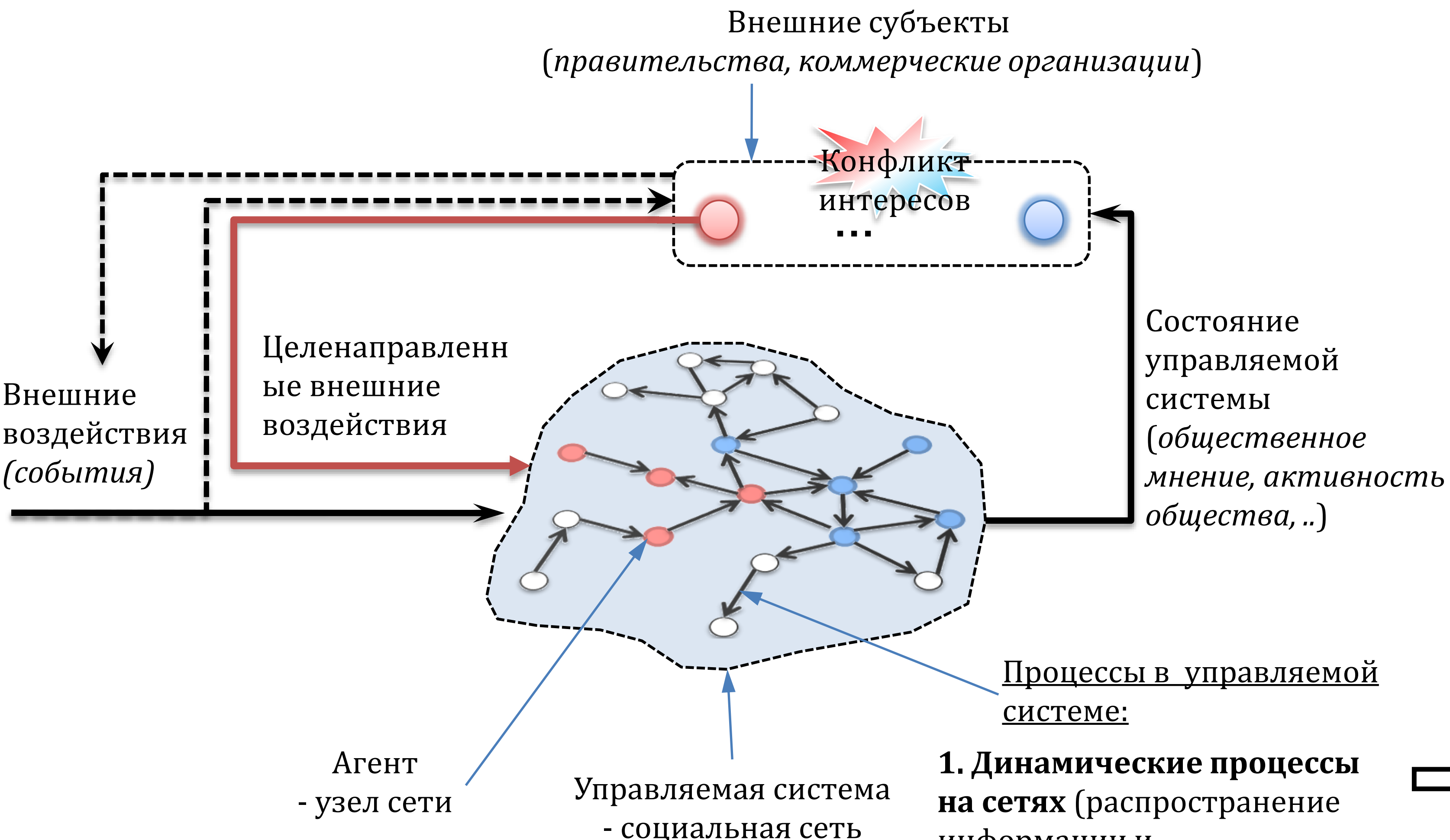
Социальные сети

- **Онлайновые социальные сети**
  - **Онлайновые сообщества**: Facebook (> 1.3 миллиардов пользователей), LinkedIn (> 380 миллионов пользователей).
  - **Коммуникация**: Skype (> 600 миллионов пользователей).
  - **Новостные и социальные медиа**: блоги (свои блоги ведут представители органов власти).
- **Мотивация исследователей**
  - Как возникают связи между пользователями социальной сети?
  - Как распространяется информация по сети? Как возникают информационные каскады? Какие пользователи являются их инициаторами?
  - Какие пользователи социальных сетей являются самыми влиятельными?
- **Мотивация практиков**: маркетинг, информационная безопасность, политика, бизнес – там, где нужно учитывать взаимодействие между пользователями.

Социальные сети и управление



Социальные сети и управление



1. Динамические процессы на сетях (распространение информации и формирование мнений)

2. Структурные процессы (формирование связей в сети)

⇨ **Модели будут представлены далее**

Общая постановка задачи управления

Объект управления описывается графом. Управление может заключаться в целенаправленном воздействии на следующие компоненты объекта управления:

– состав управляемой системы (то есть управление может заключаться в удалении или добавлении вершин);

– структуру (связи между элементами) управляемой системы (то есть управление может заключаться в удалении или добавлении дуг);

– значения параметров, соответствующих вершинам графа (значения состояний) и его дуг (значения параметров, отражающих взаимосвязи между элементами системы).

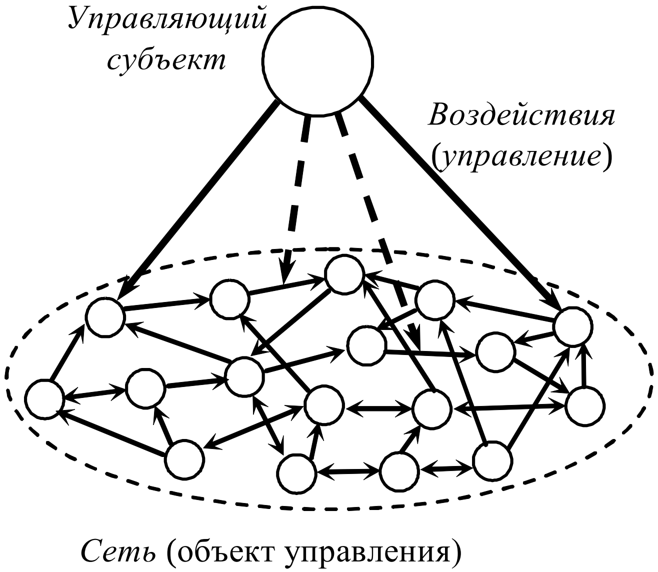
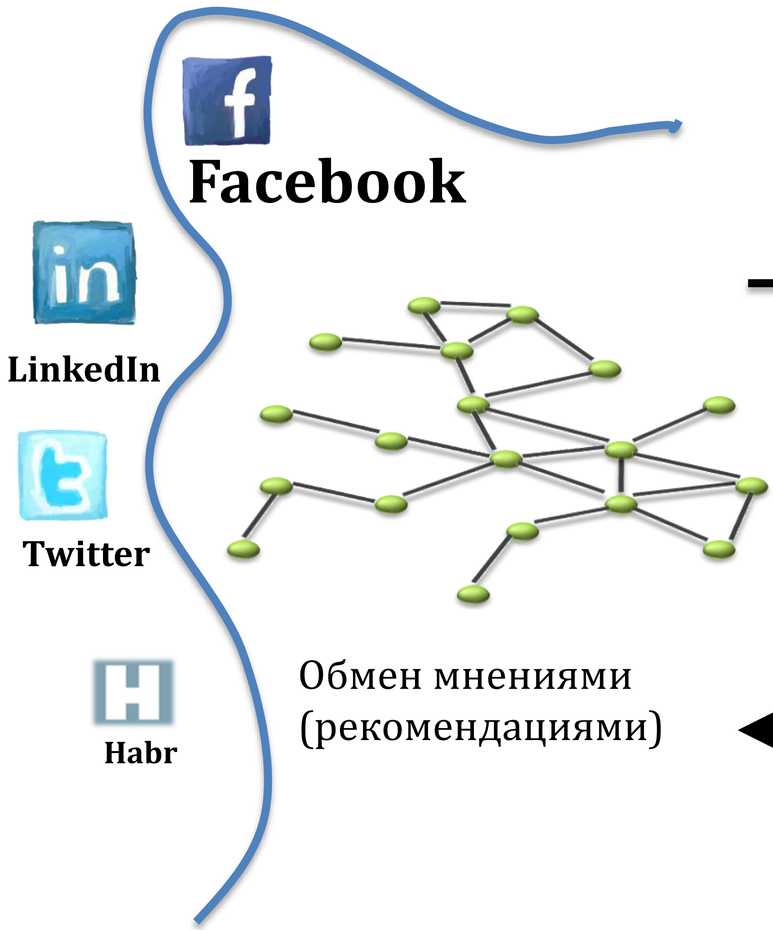


Рис. Объект управления, описываемый сетью

Пример управления в социальных сетях



Субъект управления (цель – создать требуемое мнение в сети)

- Использует модели социальных сетей и алгоритмы для поиска влиятельных пользователей в сети или для увеличения влияния уже имеющихся агентов
- Оказывает воздействие на влиятельных пользователей
- Требуемое общественное мнение

Содержание лекции

Концептуальная модель распространения информации

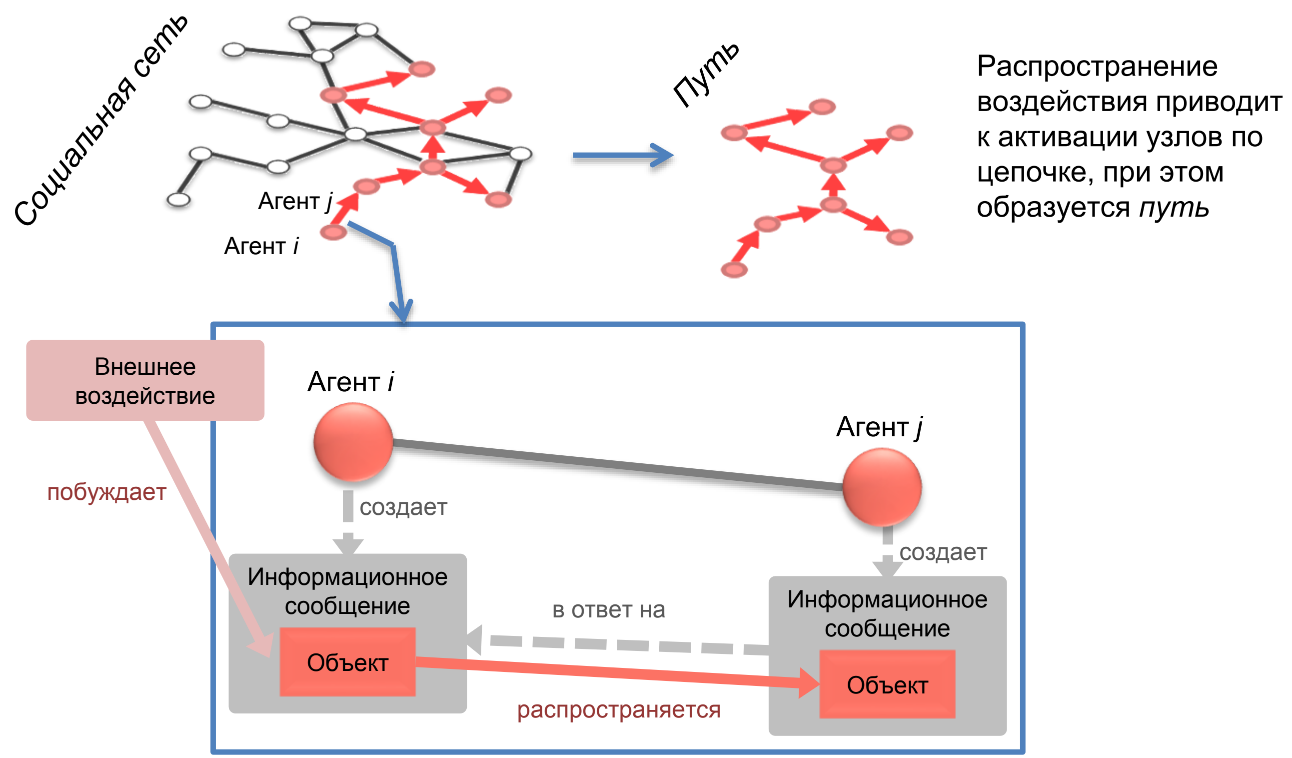
Модели распространения информации

Информационное управление

Информационное противоборство

Распространение информации

Распространение информации – процесс, посредством которого некоторый информационный объект (информация, вирус, мнение) распространяется по коммуникационным каналам во времени и в пространстве среди узлов сети



Математические модели распространения активности

Модель распространения нововведений

Динамика процесса распространения изменений моделируется S-образной (логистической) кривой

Стадии процесса

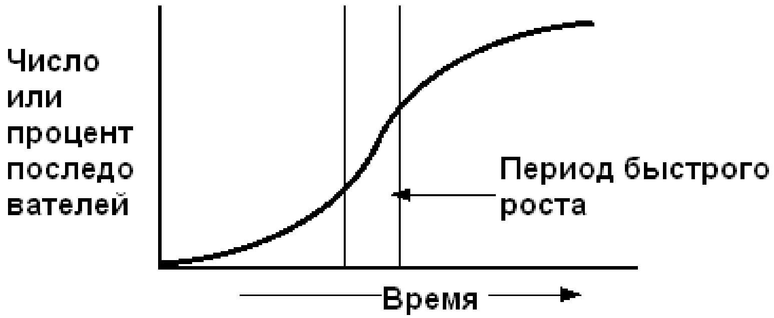
Новаторы, ранние последователи, раннее большинство, позднее большинство, поздние последователи

Стадии процесса принятия нововведения агентом

Знание, убеждение, решение, апробация, подтверждение.

Свойства нововведения

- относительные преимущества перед аналогами
- совместимость нововведения
- сложность нововведения
- простота апробации использования
- коммуникационная наблюдаемость



S-образная кривая (логистическая функция)

Кривая стадий

Модель Басса

Пусть N_0 – число агентов в сети, заданной полносвязным графом. Агент может находится в двух состояниях – активном или неактивном.

Агент переходит в активное состояние под влиянием других пользователей сети или под внешним воздействием. Динамика числа пользователей:

$$\frac{f(t)}{1-F(t)} = p + qF(t),$$

где $f(t)$ – доля пользователей, активировавшихся в момент времени t, $F(\tau) = \int_0^{\tau} f(t)dt$, p – коэффициент внешнего влияния, q – коэффициент внутреннего влияния.

Модель Басса

Пусть N_0 - число пользователей в сети,$N(t) = N_0F(t)$ - общее число пользователей, активных к моменту времени t, а $n(t)=N_0f(t)=dN(t)/dt$ - число пользователей, активировавшихся в момент времени t. Тогда уравнение динамики распространения активности:

$$\frac{dN(t)}{dt} = \left[p + \frac{qN(t)}{N_0}\right](N_0 - N(t))$$

Это уравнение решается аналитически:

$$N(t) = N_0\left[\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}}\right]$$

График функции $N(t)$ характеризуется S-образной кривой.

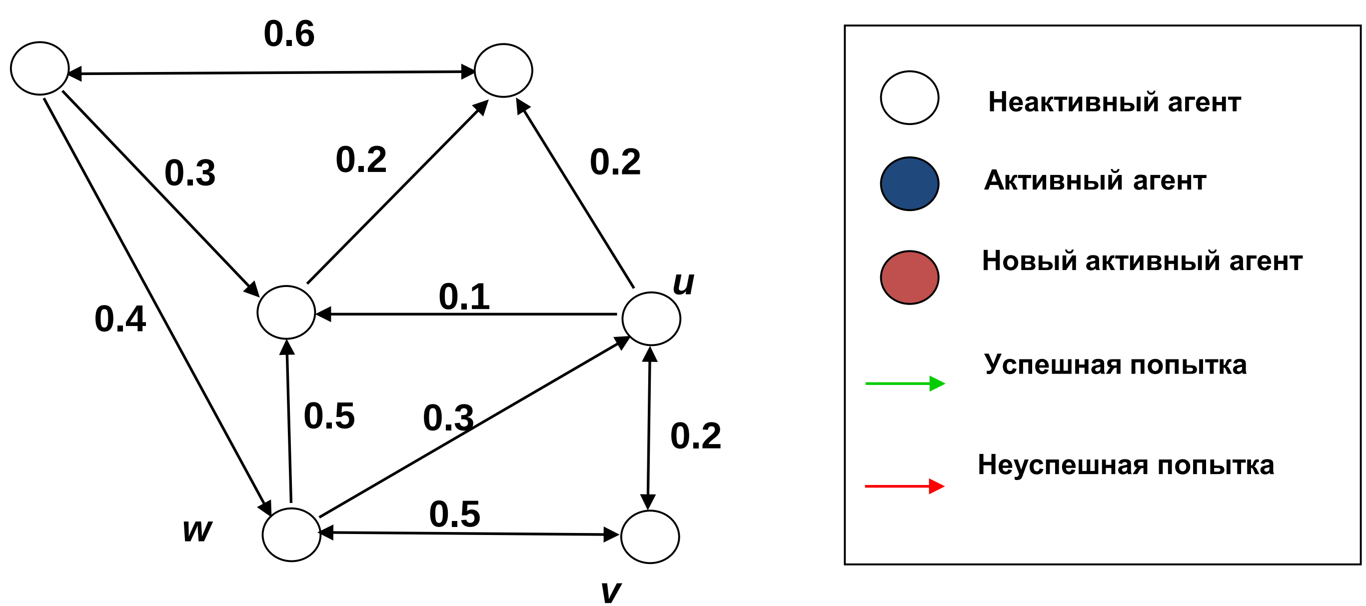
Модель независимых каскадов

Задана сеть *G(V, E).* Агент в сети может находиться в активном и неактивном состояниях, причем возможен переход только из неактивного состояния в активное.

Если агент i становится активным в некоторый момент времени, он получает шанс активировать на следующем (и только) шаге каждого из своих соседей j с вероятностью $p_{ij} \in [0, 1]$. Причем агента j могут пытаться независимо активировать и другие агенты.

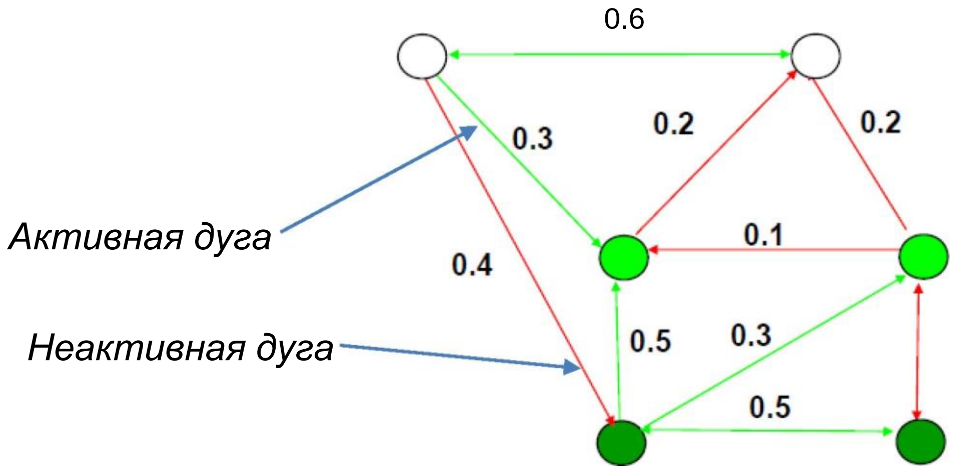
Модель независимых каскадов

Известно множество активных агентов $S_0 \subseteq V$ в момент времени $t = 0$. Для вычисления $S_\tau \subseteq V$ в момент $\tau \geq 1$ применяется итеративная процедура: на каждом шаге t ($\tau \geq t \geq 1$) сначала множество $S_t = S_{t-1}$, затем каждый агент $i \in S_{t-1} \setminus S_{t-2}$ пытается активировать еще неактивных соседей согласно правилу выше, активированные им соседи добавляются во множество S_t.



Граф активных дуг

Задан граф *G(V, E).* Каждая дуга помечается как активная или неактивная согласно определенному рандомизированному правилу. Результирующий случайный подграф, состоящий из всех узлов из *V* и всех активных дуг, назовем *графом активных дуг.*



Введем следующие обозначения:

$d_G(S, v)$ - минимальное расстояние от множества узлов *S* к узлу *v.*

$R^i_G(S)$ - множество узлов, достижимых из множества *S* за *i* шагов.

$R^i_G(S) = \{v \in V \mid d_G(S, v) \leq i\}$, в частности, $R^0_G(S) = S$.

$R_G(S)$ - множество узлов из *V*, достижимых из *S* в графе *G.*

$R^{n-1}_G(S) = R_G(S)$, $n = |V|$.

Модель активных дуг

Задан $G(V, E)$ и вероятность влияния $p(\cdot)$ *на* всех дугах. Случайный граф активных дуг G_L (или один из возможных миров) формируется посредством независимой активации каждой дуги $(u,v) \in E$ с вероятностью $p(u,v)$.

Задано начальное множество S_0, для любого шага $t \geq 1$ множество активных узлов $S_t \leftarrow R_{G_L}^t(S_0)$.

Теорема 1. *Модель независимых каскадов эквивалентна модели активных дуг с независимой активацией.*

Эквивалентность модели независимых каскадов и модели активных дуг. Доказательство (1)

Зафиксируем начальное множество S_0. Для любого $t \geq 1$ рассмотрим любую такую последовательность множеств $\boldsymbol{A_1, A_2, \ldots, A_t} \subseteq V$, что $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_t$, в которой если два последовательных множества равны, то последующие множества также равны им. Докажем, что для любой такой последовательности

$$P(S_t = A_t | S_1 = A_1, \ldots, S_{t-1} = A_{t-1}) =$$
$$P\left(R_{G_L}^t(S_0) = A_t \middle| R_{G_L}^1(S_0) = A_1, \ldots, R_{G_L}^{t-1}(S_0) = A_{t-1}\right)$$

В модели независимых каскадов.

Если задано S_0, $S_i = A_i$, ..., $S_{t-1} = A_{t-1}$, то в момент времени t множество активных узлов $S_t = A_t$, тогда и только тогда, когда:

1) для каждого узла $v \in A_t \setminus A_{t-1}$ существует активировавший его узел из $A_{t-1} \setminus A_{t-2}$;
2) не существует узла в $V \setminus A_t$, активированного каким-либо узлом из $A_{t-1} \setminus A_{t-2}$.

Поскольку все события активации независимы, то

$$P(S_t = A_t | S_1 = A_1, \ldots, S_{t-1} = A_{t-1}) =$$
$$\prod_{v \in A_t \setminus A_{t-1}} \left[1 - \prod_{u \in A_{t-1} \setminus A_{t-2}} (1 - p(u,v))\right] \prod_{v \in V \setminus A_t} \prod_{u \in A_{t-1} \setminus A_{t-2}} (1 - p(u,v))$$

Эквивалентность модели независимых каскадов и модели активных дуг. Доказательство (2)

A_1, A_2, ..., A_{t-1}, $A_t \subseteq V$, что $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_t$

В модели активных дуг с выбором независимых дуг.

Задано $R^1_{G_L}(S_0) = A_1, \ldots, R^{t-1}_{G_L}(S_0) = A_{t-1}$. Достижимое на шаге t множество $R^t_{G_L}(S_0) = A_t$ тогда и только тогда, когда:

1) каждый узел $v \in A_t \setminus A_{t-1}$ достижим за один шаг хотя бы из одного узла из $A_{t-1} \setminus A_{t-2}$;

2) не существует узла в $V \setminus A_t$ достижимого за один шаг из какого-либо узла из $A_{t-1} \setminus A_{t-2}$.

Поскольку в модели активных дуг узел u достигает узел v за один шаг с вероятностью $p(u,v)$ и каждое такое событие не зависит от другого события, то

$$P\left(R^t_{G_L}(S_0) = A_t \middle| R^1_{G_L}(S_0) = A_1, \ldots, R^{t-1}_{G_L}(S_0) = A_{t-1}\right)$$
$$= \prod_{v \in A_t \setminus A_{t-1}} \left[1 - \prod_{u \in A_{t-1} \setminus A_{t-2}} (1 - p(u,v))\right] \prod_{v \in V \setminus A_t} \prod_{u \in A_{t-1} \setminus A_{t-2}} (1 - p(u,v)) \qquad (2)$$

Уравнения (1) и (2) абсолютно одинаковы и таким образом для любого A_1, A_2, ..., $A_t \subseteq V$ выполняется

$$P(S_t = A_t | S_1 = A_1, \ldots, S_{t-1} = A_{t-1}) =$$
$$P\left(R^t_{G_L}(S_0) = A_t \middle| R^1_{G_L}(S_0) = A_1, \ldots, R^{t-1}_{G_L}(S_0) = A_{t-1}\right)$$

Свойства модели независимых каскадов

Следствие из теоремы 1. О несущественности задержек попыток активации в модели независимых каскадов

В модели независимых каскадов задержка попыток активации одного узла другим узлом не изменит распределения узлов на результирующем активном множестве $\Phi(S_0)$.

Если же разрешить узлу *u* несколько независимых попыток активации узла *v* (где $p_i(u, v)$ - вероятность *i*-ой попытки активации), то распределение узлов на результирующем активном множестве $\Phi(S_0)$ будет таким же как в модели независимых каскадов, в которой вероятность активации *v*-ого узла *u*-ым узлом будет равна $p(u, v) = 1 - \prod_i (1 - p_i(u, v))$, где $p_i(u, v)$ - вероятность *i*-ой попытки активации.

Свойства модели независимых каскадов

Понятие субмодулярности функции.

Функция множества $f\colon 2^V \to \mathbb{R}$ субмодулярна, если для любых подмножеств $S \subseteq T \subseteq V$ и любого элемента $v \in V \setminus T$:

$f(T \cup \{v\}) - f(T) \leq f(S \cup \{v\}) - f(S)$

Понятие монотонности функции.

Функция множества $f\colon 2^V \to \mathbb{R}$ монотонна, если для любых подмножеств $S \subseteq T \subseteq V$ выполняется $f(S) \leq f(T)$.

Теорема 2. Функция распространения влияния $\sigma(\cdot)$ в модели независимых каскадов является монотонной и субмодулярной.

Свойства модели независимых каскадов

Доказательство теоремы 2.

Задан *G*(*V*, *E*), задано Г – множество всех возможных графов активных дуг из *G*. Пусть G_L - случайный граф активных дуг, а P(G_L) - вероятность реализации графа G_L.

Для модели независимых каскадов (ICM эквивалентна модели активных дуг с выбором независимых дуг) математическое ожидание финального количества активных узлов:

$$\sigma(S_0) = \sum_{G_L \in \Gamma} P(G_L) \left| R_{G_L}(S_0) \right|$$

Известно, что линейная комбинация монотонных (субмодулярных) функций с неотрицательными коэффициентами также монотонна (субмодулярна). Следовательно, для доказательства того, что $\sigma(S_0)$ монотонна (субмодулярна), достаточно показать, что для любого графа активных дуг G_L – функция $\left|R_{G_L}(\cdot)\right|$монотонна (субмодулярна).

Свойства модели независимых каскадов

Доказательство теоремы 2 (продолжение)

Монотонность. Монотнонность функции очевидна.

Субмодулярность. По определению субмодулярности достаточно показать, что для любых двух подмножеств $S \subseteq T \subseteq V$ и любого узла $v \in V \setminus T$:

$$R_{G_L}(T \cup \{v\}) \setminus R_{G_L}(T) \subseteq R_{G_L}(S \cup \{v\}) \setminus R_{G_L}(S)$$

Любой узел $u \in R_{G_L}(T \cup \{v\}) \setminus R_{G_L}(T)$ достижим из $T \cup \{v\}$, но не достижим из T в графе G_L. Следовательно, u достижим из v. Поскольку $S \subseteq T$, то должен быть такой случай, что u не достижим из S, но достижим из $S \cup \{v\}$, следовательно $u \in R_{G_L}(S \cup \{v\}) \setminus R_{G_L}(S)$.

Таким образом функция $|R_{G_L}(\cdot)|$ субмодулярна для всех G_L, следовательно субмодулярна и функция $\sigma(\cdot)$

.

Линейная модель с порогами

В линейной модели с порогами *i*-ый агент - узел социальной сети (вершина графа) испытывает влияние w_{ji} каждого своего *j-го* соседа в сети. Сумма влияний соседей на агента не превосходит единицы:

$$\sum_{j \in N_{in}(i)} w_{ji} \leq 1$$

где $N_{in}(i)$ - соседи *i* -ого агента.

Агент *i* имеет некоторый порог ϕ_i, который изначально независимо и равномерно случайно выбирается из отрезка [0, 1].

Агент в сети может находиться в активном и неактивном состояниях, причем возможен переход только из неактивного состояния в активное (обратный переход не допускается).

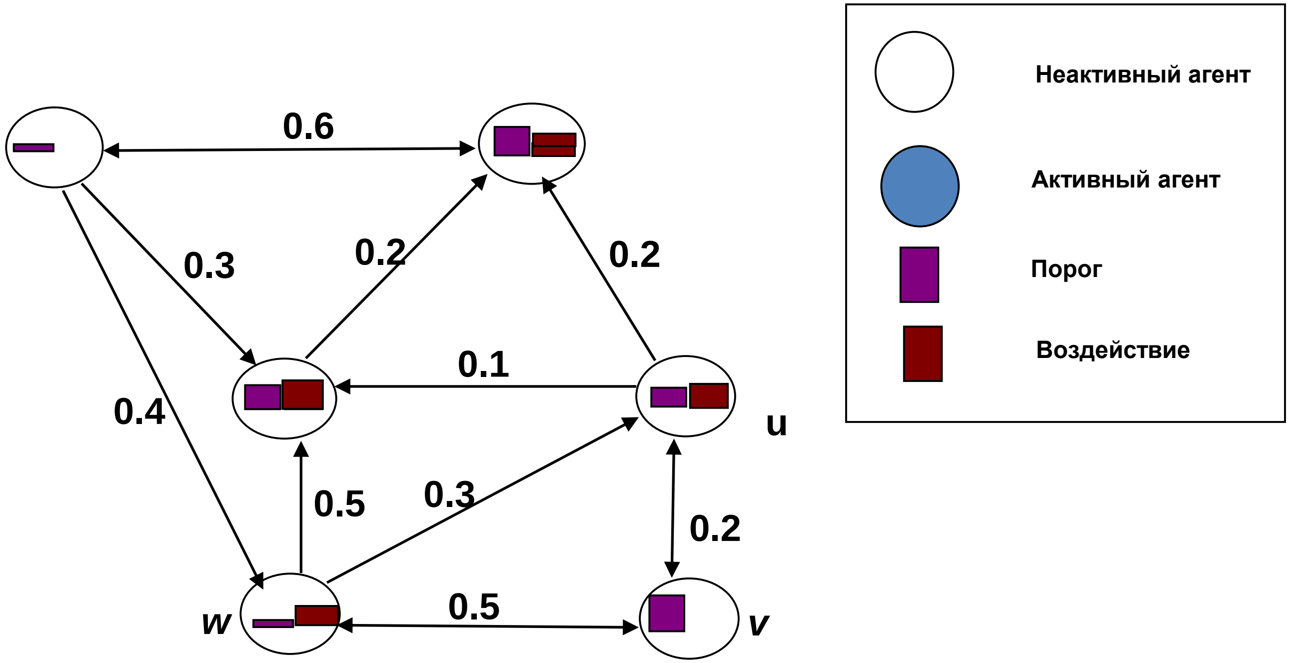
Линейная модель с порогами

Пусть в момент времени $t = 0$ известно начальное множество активных агентов $S_0 \subseteq V$, необходимо определить в любой другой момент времени $t \geq 1$ множество $S_t \subseteq V$. Для этого используется следующая итеративная процедура: в каждый момент $t \geq 1$ во множество S_t добавляется еще неактивный агент $i \in V \setminus S_{t-1}$, для которого:

$$\sum_{j \in N_{in}(i) \cap S_{t-1}} w_{ji} \geq \phi_i$$

(т.е. агент i становится активным в момент времени t). Содержательно порог ϕ_i можно трактовать как подверженность агента влиянию своих соседей.

Линейная модель с порогами. Пример.



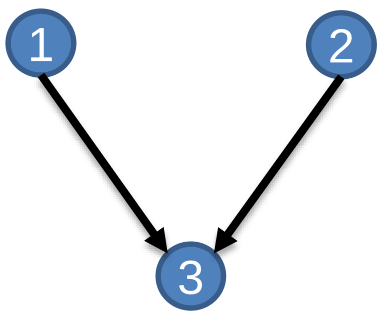
Свойства линейной модели с порогами

- В линейной модели с порогами распределение результирующего множества активных агентов $\Phi(S_0)$ не изменится, если воздействие активированного агента на своих соседей будет отложено на более поздний момент времени.

- Функция распространения влияния $\sigma(\cdot)$ в линейной модели с порогами является монотонной и субмодулярной.

Эквивалентность ICM и LTM

- В общем случае ICM и LTM не эквивалентны.
  - Пример: граф из трех узлов {1, 2, 3}, параметры ICM – *p*(1,3) и *p*(2,3), параметры LTM - *w*(1,3) и *w*(2,3).



Предположим, что ICM и LTM эквивалентны, тогда:
— если S_0 = {1}, то вероятность активации 3 узла *p*(1,3)= *w*(1,3);
— если S_0 = {2}, то вероятность активации 3 узла *p*(2,3)= *w*(2,3);
— если S_0 = {1, 2}, то вероятность активации 3 узла в ICM 1-(1-*p*(1,3))(1-*p*(2,3)), а в LTM - *w*(1,3)+*w*(2,3).
Т.е. в общем случае равенство таких вероятностей не выполняется

- Однако обобщенные ICM и LTM являются эквивалентными, т.е. для любой обобщенной ICM с заданными параметрами найдется эквивалентная LTM и наоборот.

Обобщение модели независимых каскадов

Функция активации. Вероятность $p_v(u, S)$ того, что агент u активирует агента v, зависит от множества S агентов, уже безуспешно пытавшихся активировать агента v ($u \notin S, u \in N_{in}(v), S \subseteq N_{in}(v)$).

Ограничение на функцию активации p_v: если соседи $u_1, \ldots, u_l$ пытаются активировать агента v, то вероятность того, что v станет активным после l попыток, не должна зависеть от порядка попыток активации.

Более формально: пусть $u_1, \ldots, u_l$ и $u'_1, \ldots, u'_l$ две перестановки S, и $S_i = \{u_1, \ldots, u_{i-1}\}$, $S'_i = \{u'_1, \ldots, u'_{i-1}\}$, тогда независимость от порядка означает

$$\prod_{i=1}^{l}(1 - p_v(u_i, S_i)) = \prod_{i=1}^{l}(1 - p_v(u'_i, S'_i))$$

Модель независимых каскадов является частным случаем обобщенной модели, где функция активации $p_v(u, S)$ является константой $p(u, v)$, не зависящей от S.

Обобщение линейной модели с порогами

Функция активации. Каждый агент i имеет следующую функцию активации

$f_i : S \subseteq N_{in}(i) \rightarrow [0, 1]$,

где $N_{in}(i)$ – множество соседей, влияющих на агента i, $f_i(\varnothing) = 0$.

Правило активации. Агент i становится активным в некоторый момент времени t, если $f_i(S) \geq \varphi_i$, где φ_i - порог, выбираемый из равномерного распределения (от 0 до 1), а S – множество соседей i, активных в момент t-1.

Линейная пороговая модель является частным случаем обобщенной модели с порогами, где

$$f_i(S) = \sum_{j \in S} w_{ij} \leq 1$$

Эквивалентность обобщенных моделей

Связь между $f_v(S)$ и $p_v(u, S)$.

В обобщенной пороговой модели вероятность того, что новый активный сосед *u* сможет активировать *v*, если агенты из *S* ($u \notin S$) уже безуспешно пытались активировать *v*:

$$p_v(u, S) = \frac{f_v(S \cup \{u\}) - f_v(S)}{1 - f_v(S)} \qquad (1)$$

В обобщенной каскадной модели рассмотрим узел *v* и множество его соседей $S = \{u_1, \ldots, u_k\}$. Предположим, что узлы в *S* пытаются активировать *v* в порядке $u_1, \ldots, u_k$, пусть $S_i = \{u_1, \ldots, u_i\}$. Тогда вероятность активации *v*:

$$f_v(S) = 1 - \prod_{i=1}^{k} (1 - p_v(u_i, S_{i-1})) \qquad (2)$$

Теорема об эквивалентности. Для любой обобщенной каскадной модели с функциями активации $p_v(u, S)$ можно построить эквивалентную обобщенную пороговую модель с $f_v(S)$, заданными уравнением (1). Аналогично, для любой обобщенной пороговой модели с функциями активации $f_v(S)$ можно построить эквивалентную обобщенную каскадную с $p_v(u, S)$, заданными уравнением (2).

Свойства обобщенных моделей

Теорема о монотонности. Функция распространения влияния $\sigma(\cdot)$ для любой обобщенной модели монотонна.

Теорема о субмодулярности. Функция распространения влияния $\sigma(\cdot)$ для пороговой обобщенной модели субмодулярна, если для всех агентов $v \in V$ функции активации $f_v(\cdot)$ субмодулярны.

Теорема о распределениях. Обобщенная пороговая модель с пороговыми функциями $f_v(S)$, в которой для любых $v \in V$ порог θ_v выбирается случайно согласно распределению $F_v(\cdot)$, эквивалентна обобщенной пороговой модели с пороговыми функциями $f'_v(S)$, в которой $f_v(S)= F_v(f_v(S))$ и пороги выбраны случайно согласно равномерному распределению.

Обучение моделей распространения информации

- Задача вывода:
  - Известна структура социальной сети.
  - Известно то, как распространялась информация в прошлом.
  - Необходимо получить значения влияния одних пользователей на других

Обучение моделей распространения информации: модель независимых каскадов

- $D(0), D(1), \ldots, D(t)$ – узлы, активированные в момент времени $\tau = 0,1,..t$
- $P_w(t+1) = 1 - \prod_{v \in N^{in}(w) \cap D(t)} (1 - p_{vw})$ – вероятность активации w
- Необходимо найти $\theta = \{p_{vw}\}$, максимизирующее

$$\boldsymbol{L}(\boldsymbol{\theta}; \boldsymbol{D}) = \left(\prod_{t=0}^{T-1} \prod_{w \in D(t+1)} P_w(t+1) \right) \cdot \left(\prod_{t=0}^{T-1} \prod_{v \in D(t)} \prod_{w \in N^{out}(v) \setminus \cup_{\tau \le t} D(\tau)} (1 - p_{vw}) \right)$$

Недостатки:

— весьма затратные расчеты,

— предполагается, что веса влияний не меняются во времени.

Задачи управления распространением информации

- **Минимизация распространения**

В сети распространяется нежелательная информация или вирус (рис. А). Состояние каких узлов необходимо отслеживать, чтобы эффективно выявлять возможные «эпидемии»?

- **Максимизация распространения**

Новый продукт появился на рынке (рис. Б). Кому нужно раздать свободные образцы для того, чтобы максимизировать продажи продукта?

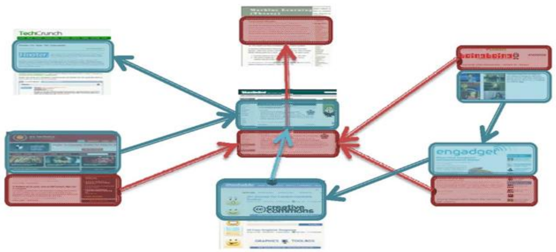


Рис. А. Пример «кросспостинга» в блогах*

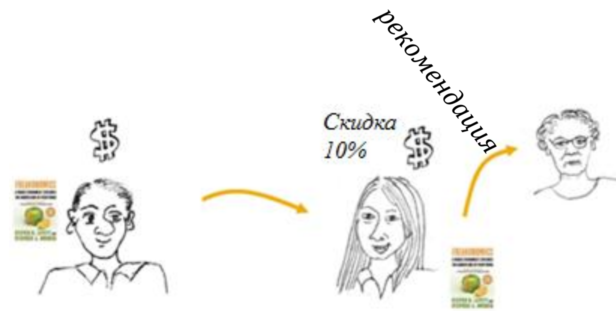


Рис. Б. Пример вирусного маркетинга*