Information extraction methods from network sources

Lecture 1

Structure of The World Wide Web

Prykhodko Tatyana

Содержание

What Is the Web?

What Is the Web

How do you think:

Is there some difference between WWW and Internet?

What Is the Web

The **World Wide Web** (**WWW**) is

- an ***open source information space*** where
- documents and other web resources are **identified by URLs**,
- ***interlinked by hypertext links***,
- and ***can be accessed via the Internet***.

It has become known simply as **the Web**.

- The Web != Internet
- The World Wide Web is **an application of the Internet**

WEB structure and content

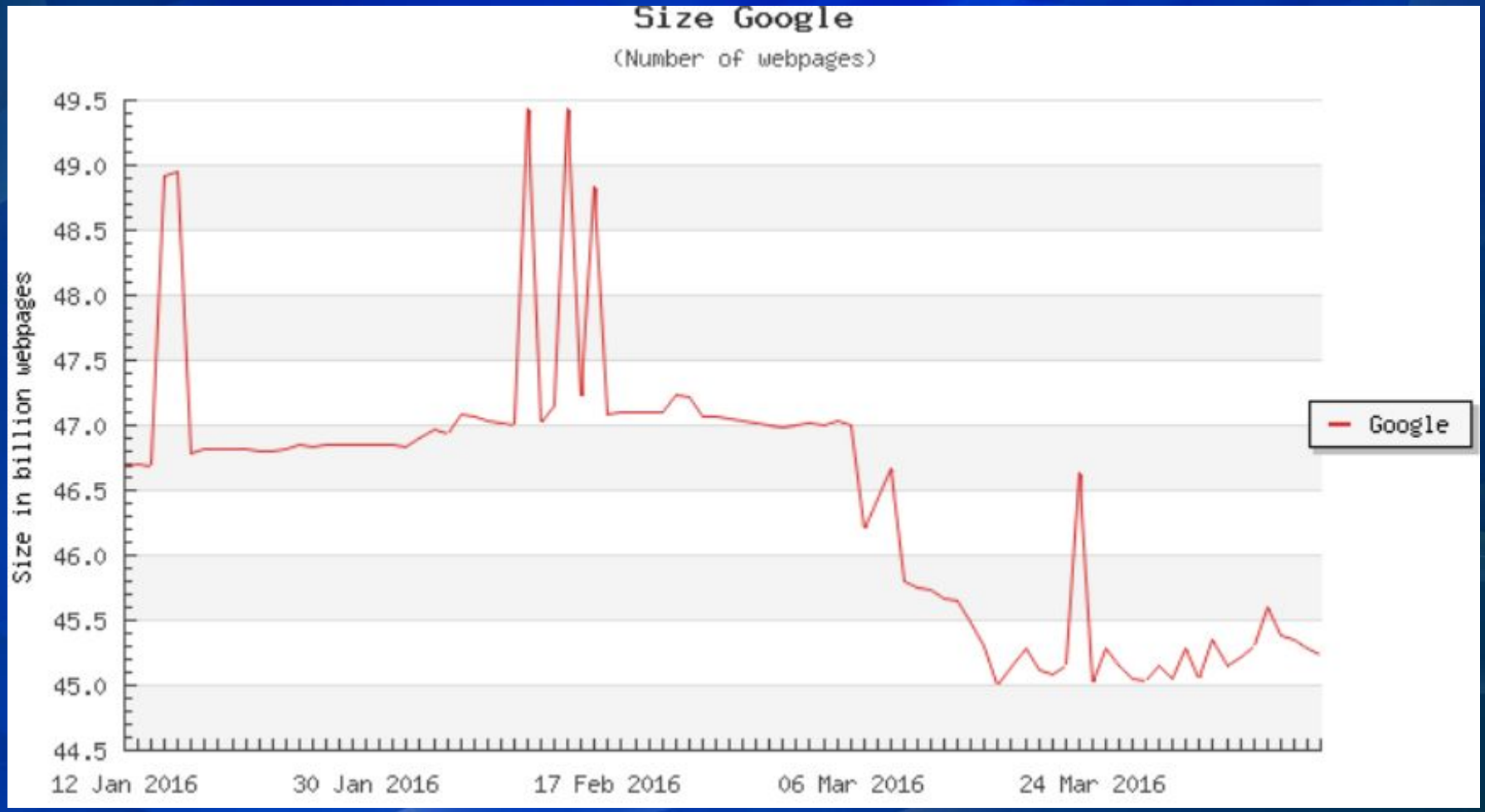
- The basic units - connected (nodes) are pieces of information
- The edges symbolize some kind of connection between them
- Share a lot of the ideas

Size of the Web

- Number of pages
  - Technically, infinite (because of dynamically generated content)
  - Much duplication (30-40%)
  - Best estimate of “unique” static HTML pages comes from search engine claims
    - The Indexed Web contains **at least 4.84 billion pages** (10^9) (Monday, 18 January, 2016).
    - Google recently announced that their index contains 1 trillion pages (10^{12})
      - How to explain the discrepancy?

Size of the Web

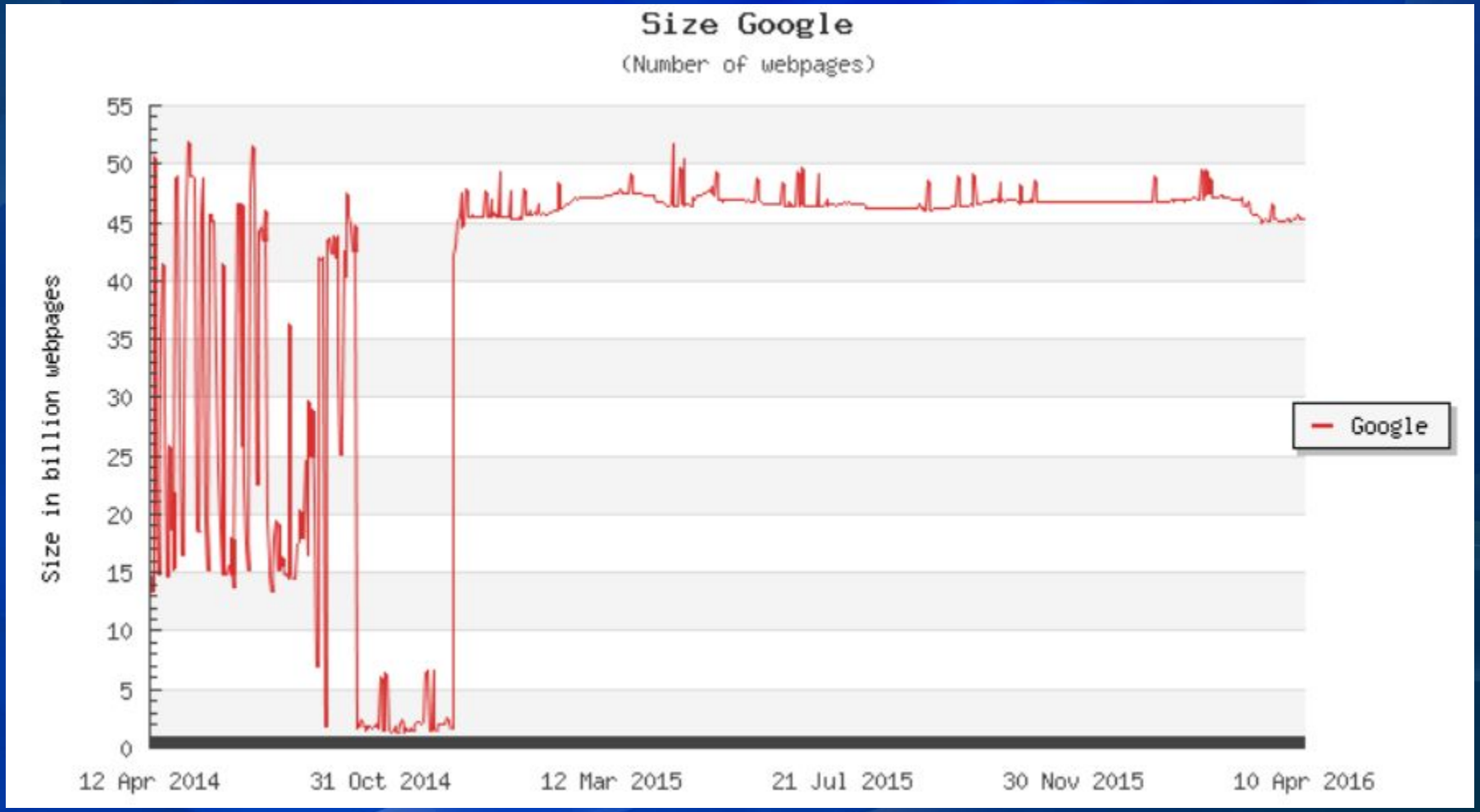
The size of the World Wide Web: Estimated size of Google's index for last 3 month



http://www.worldwidewebsize.com/

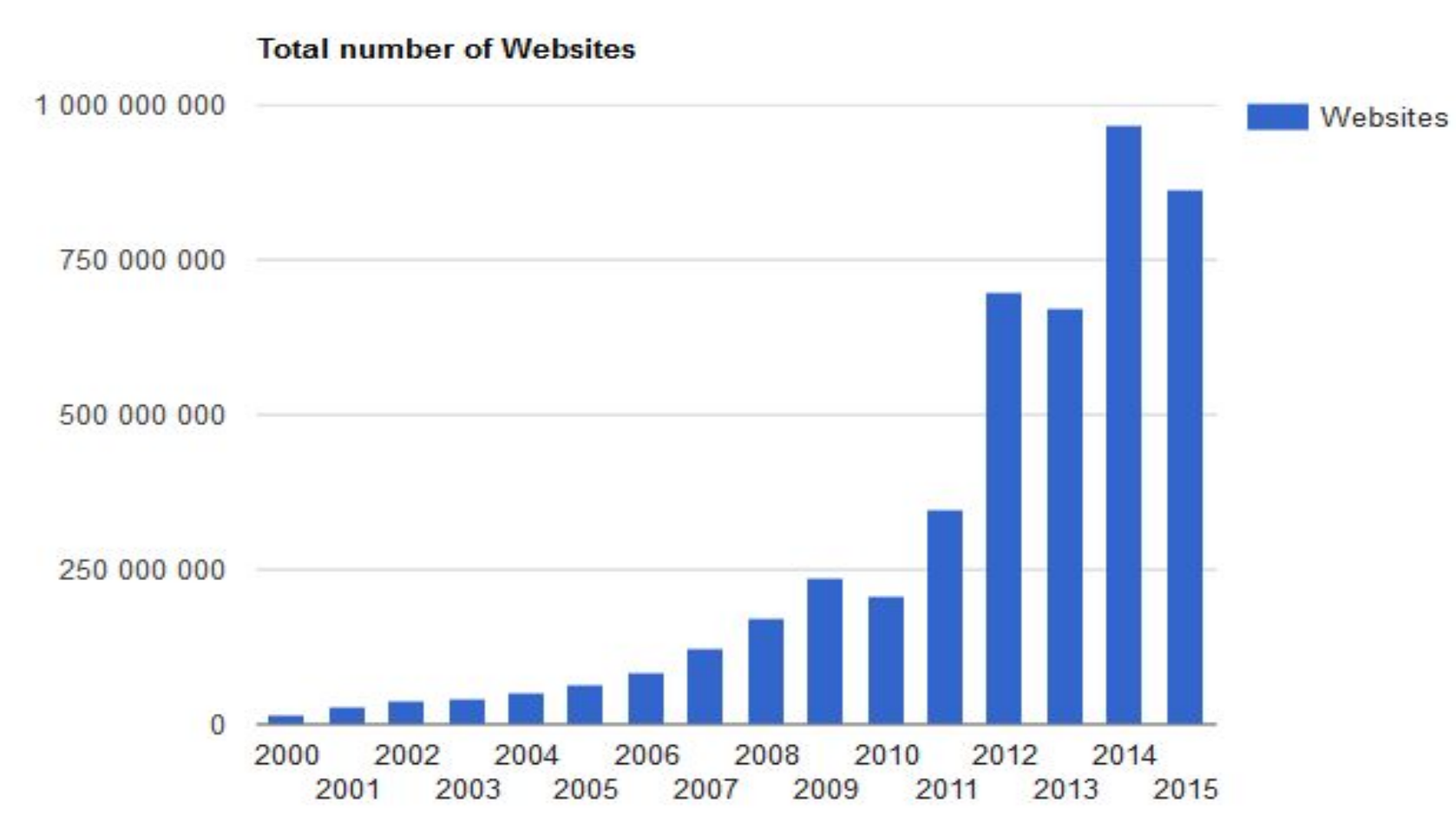
Size of the Web

The size of the World Wide Web: Estimated size of Google's index for 2 years



http://www.worldwidewebsize.com/

Size of the Web



http://www.internetlivestats.com/total-number-of-websites/

Size of the Web

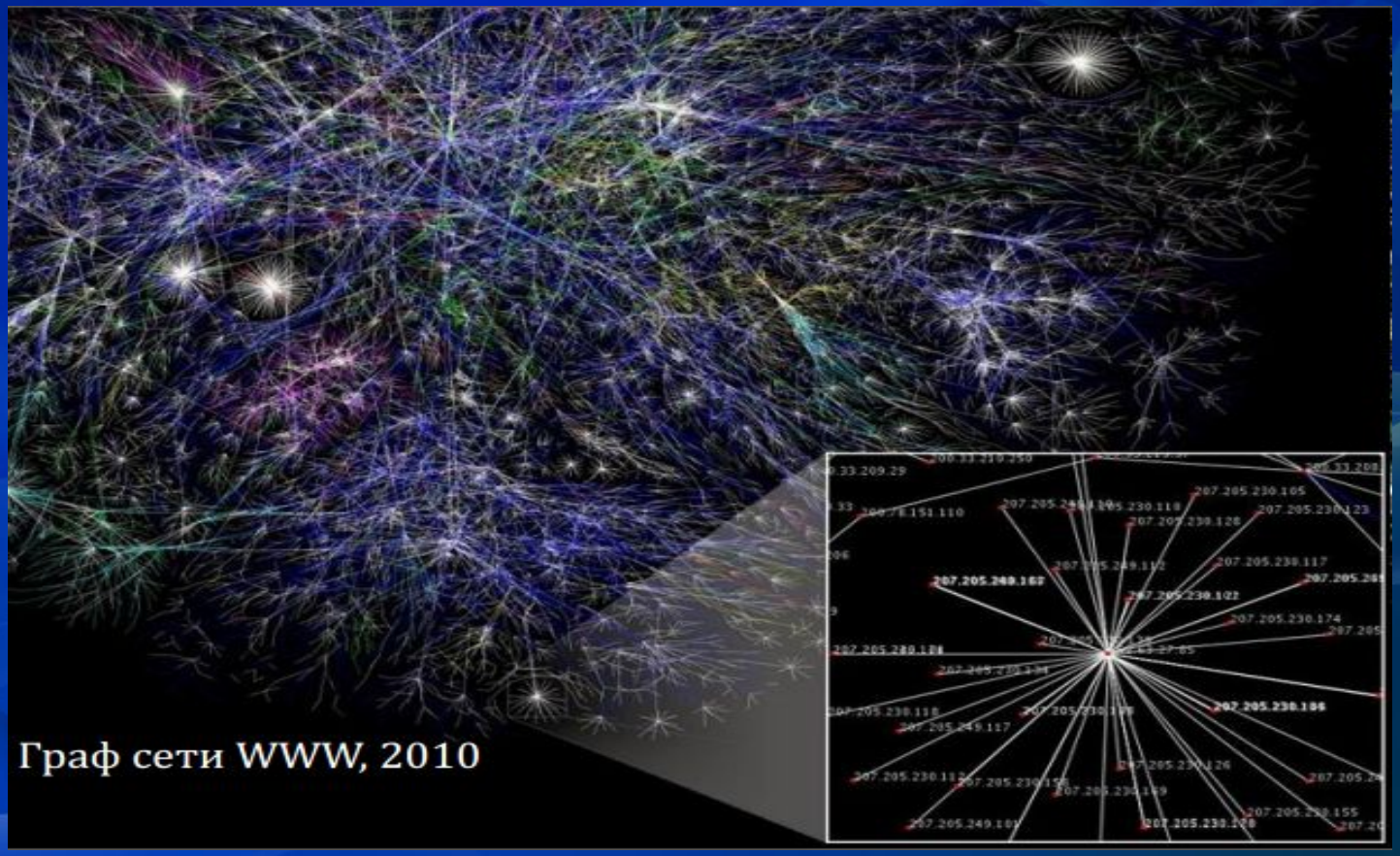
- **How Many People Would It Take Memorise The Internet?**

- If the web is equivalent to 4 zettabytes (or 4,000,000,000,000,000,000,000 bytes) Sextillion/Trilliard = 10^{21}
- and the memory capacity of a person from 10^{12} (terabyte) up to 2.5 *10^{15} (petabyte)
- then currently, in 2013, it would take around 2 *10^{6} people to store the Internet – in their heads.

http://www.infoniac.ru/news/Skol-ko-megabait-vmeshaet-chelovecheskii-mozg.html
https://en.wikipedia.org/wiki/Zettabyte

Size of the Web

- **To appreciate the task of WEB search and analysis we need some structure to view the WEB.**



The web as a graph

- Pages = nodes, hyperlinks = edges
  - Ignore content
  - Directed graph
- High linkage
  - 10-20 links/page on average
  - Power-law degree distribution

What can the graph tell us?

- Distinguish “important” pages from unimportant ones
  - Page rank
- Discover communities of related pages
  - Hubs and Authorities
- Detect web spam
  - Trust rank

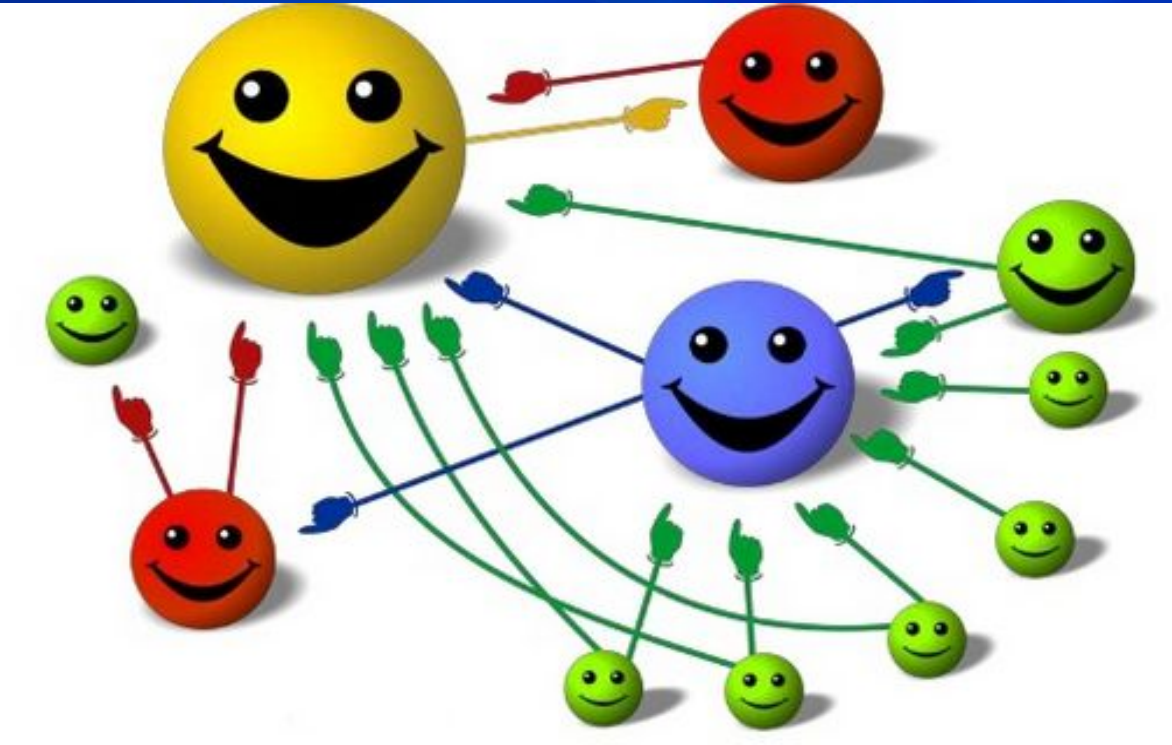
What can the graph tell us?

PageRank

- ***PageRank*** is an algorithm used by Google Search ***to rank websites in their search engine results***.
- PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of ***measuring the importance of website pages***. According to Google:
  - PageRank works by counting the number and quality of ***links to a page*** to determine a rough estimate of how important the website is. The underlying assumption is that ***more important websites are likely to receive more links from other websites.***

What can the graph tell us?

PageRank

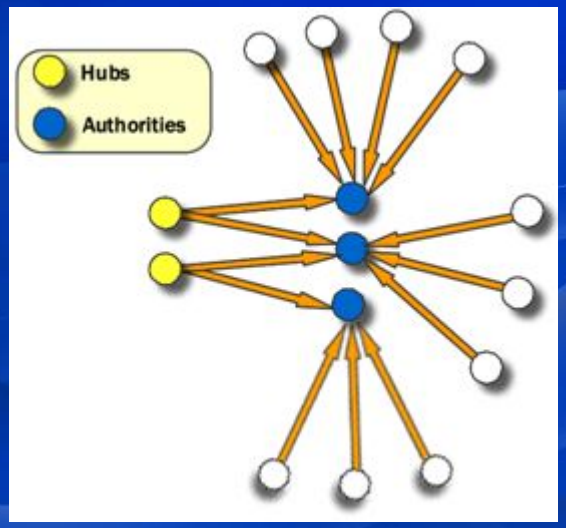


- The picture is illustrating the basic principle of PageRank.
- The size of each face is proportional to the total size of the other faces which are pointing to it.

What can the graph tell us?

Hub and Authority

- Entities that many other entities point to are called Authorities. Relationships are directional—they point from one entity to another. If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  - Is a knowledge or organizational authority within a domain.
  - Acts as definitive source of information.

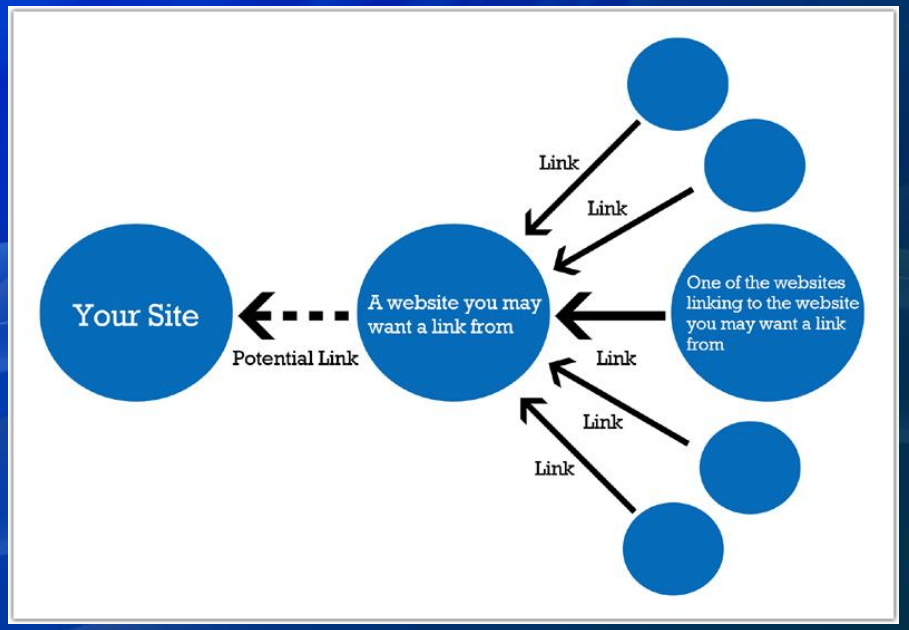


- Hubs are entities that point to a relatively large number of authorities.
- They are essentially the mutually reinforcing analogues to authorities.
- Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.

What can the graph tell us?
TrustRank

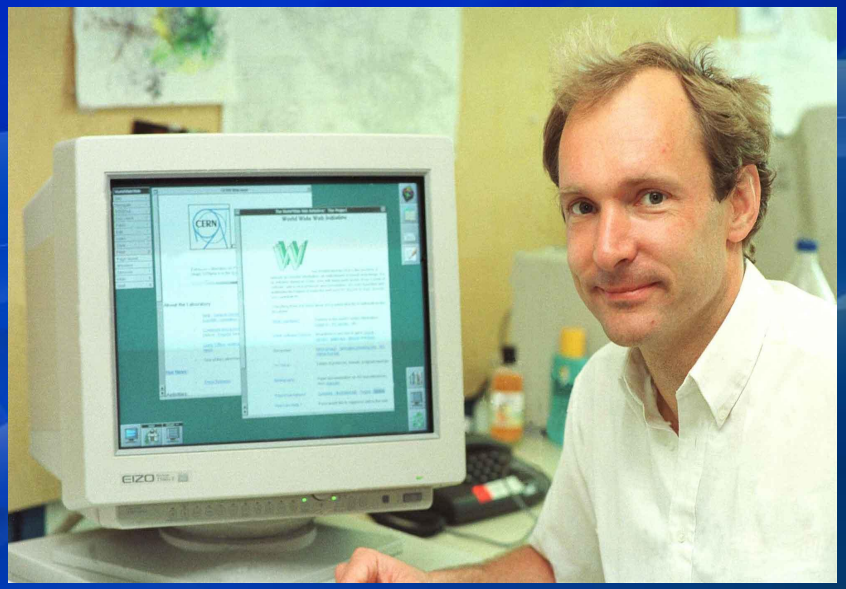
- **TrustRank** is a link analysis technique described by researchers of Stanford University and of Yahoo!. The technique is used for semi-automatic separation of useful webpages from spam.

- The starting point of the algorithm is the selection of good (trusted) pages by hand. These pages are the sources of trust.
- Trust can be transferred to other page by linking to them. Trust is propagating in the same way as PageRank.



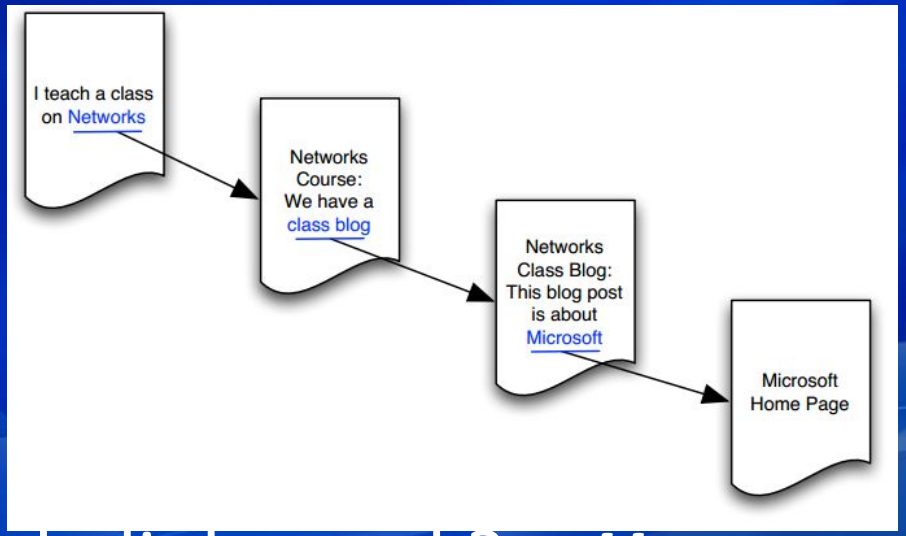
Back to the web

- Created by Tim Burners-Lee
- A research project in 1989-1991 at CERN
- An application of the Internet
- Two basic features:
  - Make documents on your computer publically accessible
  - Easily access these documents using a browser



The web as a network

- The nodes are documents (pages)
- The edges are links



- How do links work? - ***Hypertext!***

Hypertext

(The coolest thing about the web)

Different ways to manage information

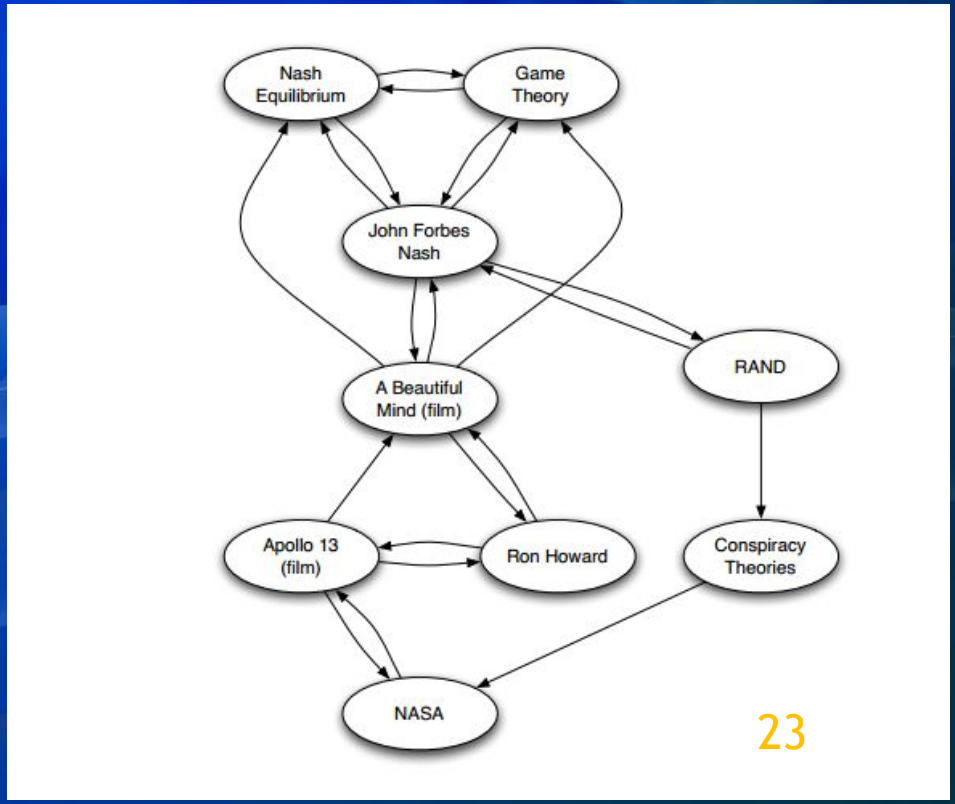
- Alphabetically
- Hierarchy (like folders)
- Classification systems

- All of these have one thing in common:

Linear

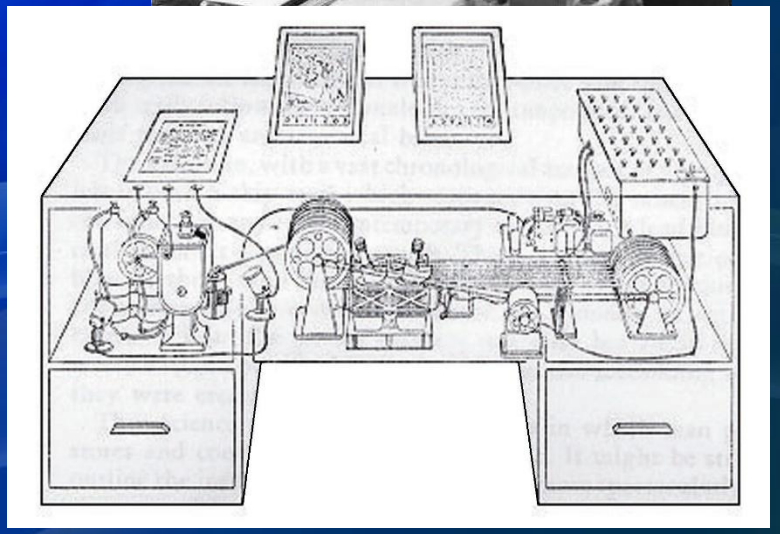
Earlier non linear connections

- Academic references
  - (also in legal decisions and patents)
- Relevant to the web?
- Cross-reference encyclopedia



Memex

- Vannevar Bush, 1945
  Article: “As We May Think”
- Our memory is not linear.
- Hypothetical model – the Memex
- Inspired the idea of hypertext



- An associative way to organize information

Changes in the web over time

Static pages >> Query (dynamic) pages

- In the early days – static pages of contact
- Today?
- More and more *transactional* actions, which create ***query pages***

Importance of static pages

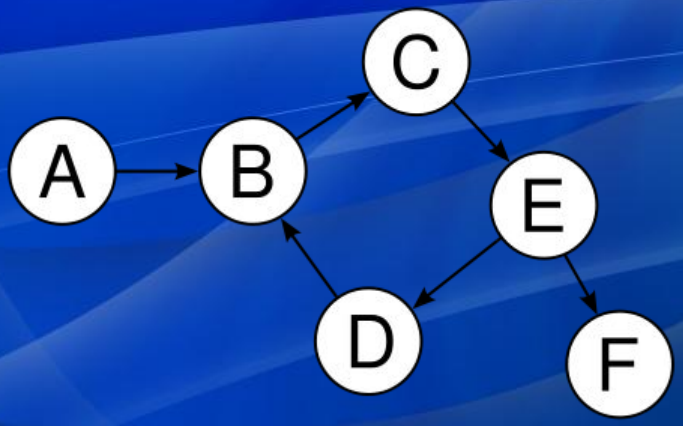
- “The Backbone of the Internet”
- Reliable over time
- Include most links
- Navigational vs. transactional
- Our focus when thinking about WEB structure

The web as a directed graph

- Viewing social and economic networks in terms of their graph structures provides significant insights, and the same is true for information networks such as the Web.
- When we view the Web as a graph, it allows us
  - to better understand the ***logical relationships expressed by its links***;
  - to break its structure into smaller, cohesive units;
  - and— to identify important pages as a step in organizing the results of Web searches.

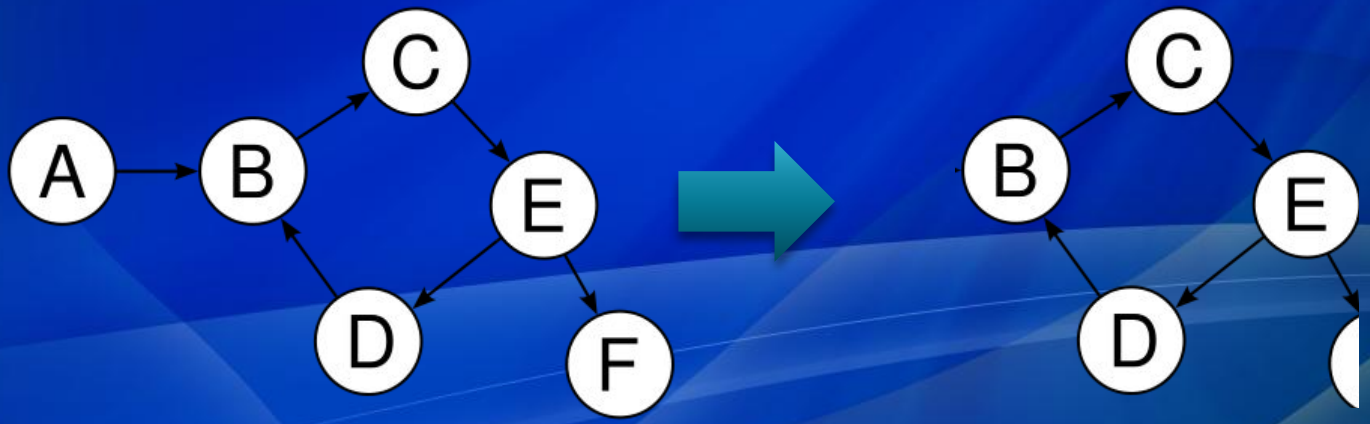
What is a path in a directed graph?

- "A *Path* from node A to a node F in a directed graph is a sequence of nodes, beginning with A and ending with F, with the property that each consecutive pair of nodes in the sequence is connected by an edge pointing in the forward direction"



What is Strong Connectivity in a directed graph?

- "A directed graph is *Strongly connected* if there is a path from every node to every other node"



The Concept of Reachability

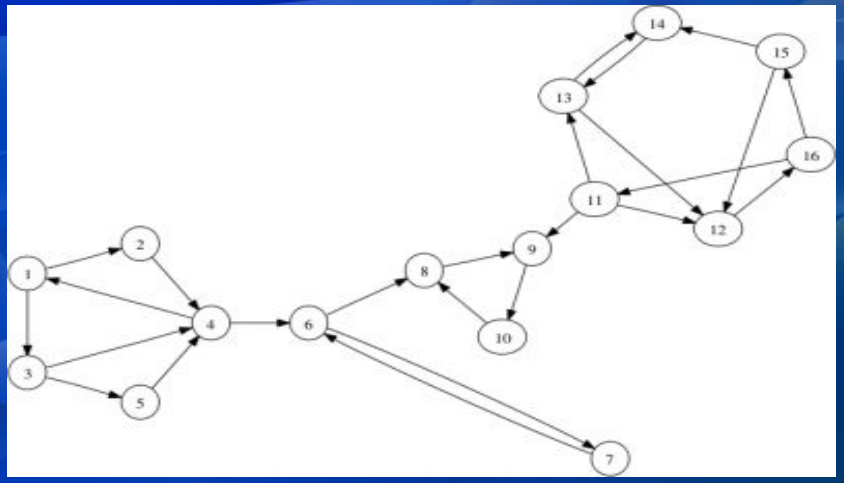
- Since connectivity does not describe all of the connections in a graph, we need another concept – Reachability
- Reachability describes the nodes that are reachable from a certain node or vice versa
- How do we check this?

Strongly connected components

- Parts of a graph that have strong connectivity
- In other words – a group of nodes in which each node is reachable from all other nodes.
- Formal:

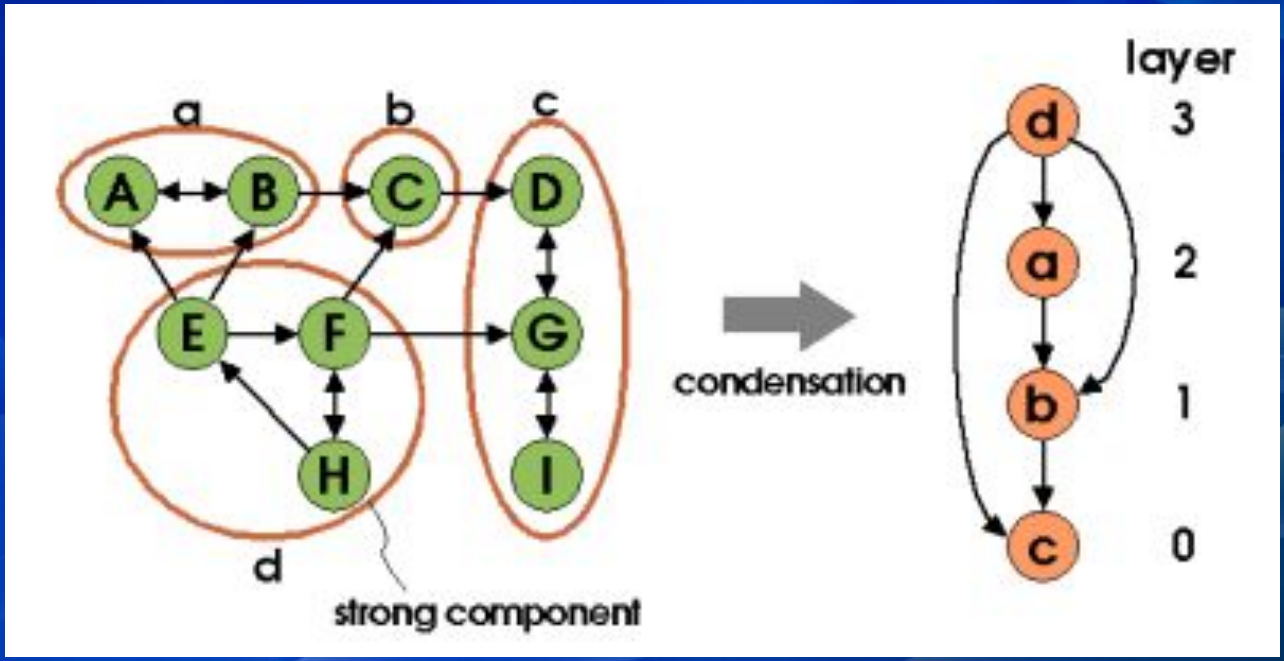
*We say that a **strongly connected component (SCC)** in a directed graph is a subset of the nodes such that:*

(a) every node in the subset has a path to every other; and
(b) the subset is not part of some larger set with the property that every node can reach every other.



How does all that help us understand the web?

- We can map reachability using the super-graph



The Bow Tie Structure

History of Bow Tie model

- Created in 1999 by Andrei Broder and his colleagues from IBM, Compaq and AltaVista
- Used data from biggest search engine back then – AltaVista.
- Afterwards – reevaluated many times

Web graph structure

- Web may be considered to have five major components
  - ***Central core*** – strongly connected component (SCC) – pages that can reach one another along directed links - about 30% of the Web
  - ***IN group*** – can reach SCC but cannot be reached from it - about 20%
  - ***OUT group*** – can be reached from SCC but cannot reach it - about 20%

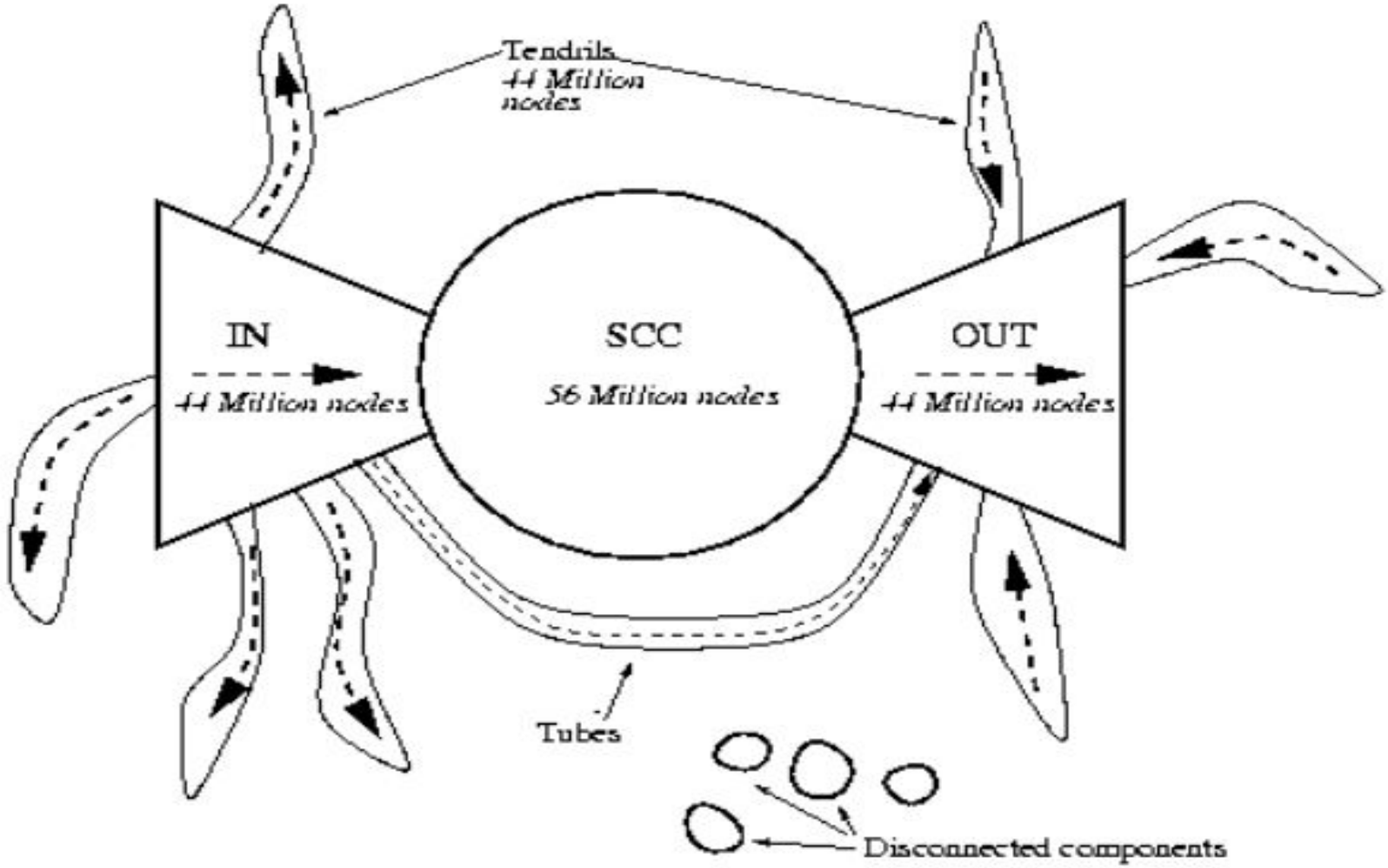
Web graph structure

- ***Tendrils*** – cannot reach SCC and cannot be reached by it - about 20%

- ***Unconnected*** – about 10%

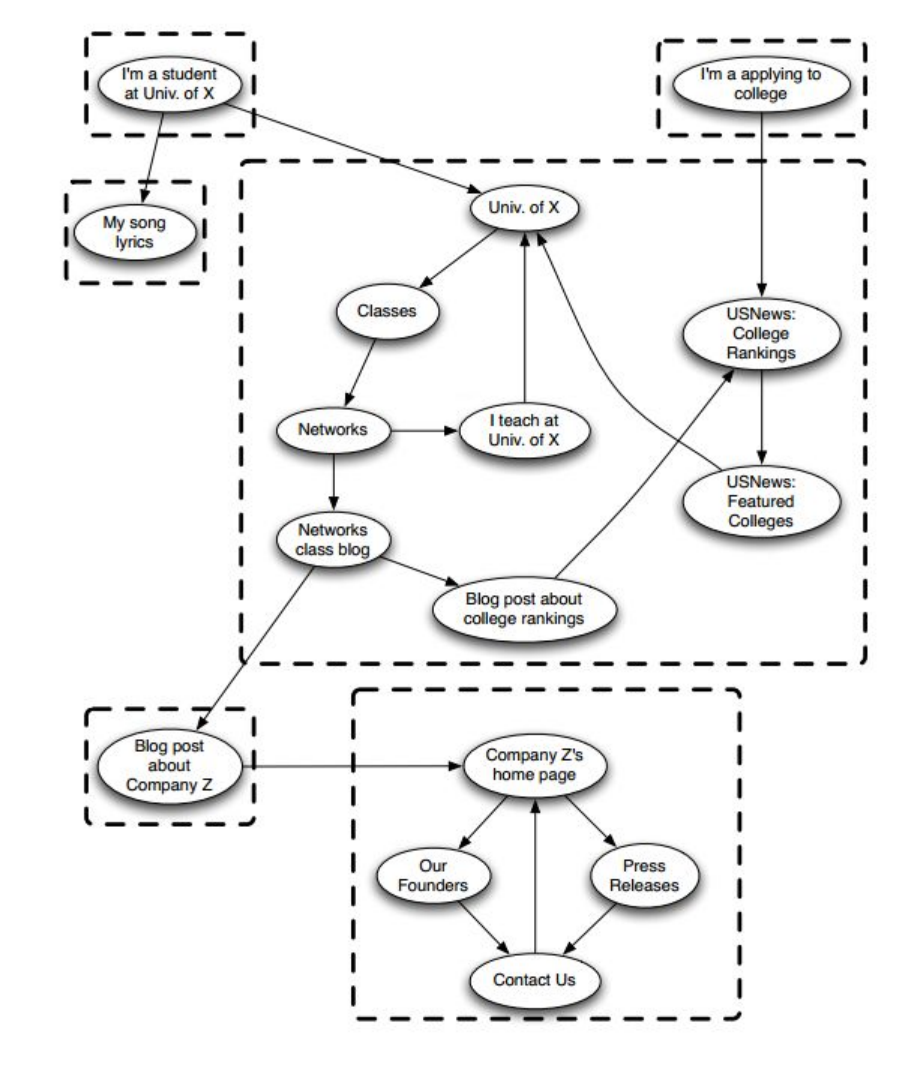
- The Web is hierarchical in nature.
- The Web has a strong locality feature. Almost two thirds of all links are to sites within the enterprise domain. Only one-third of the links are external. Higher percentage of external links are broken.
- The distance between local links tends to be quite small.

The bow tie structure



Different kinds of nodes

- In the SCC
- In the “inbound” part
- In the “outbound” part
- Tendrils
- Disconnected nodes



Некоторые дополнительные факты

- Итак, в рамках общей задачи определения структуры связей между отдельными веб-страницами было выявлено:
  - центральное ядро (28 % веб-страниц) — зона сильной связности сети *(Strongly Connected Component, SCC);*
  - «отправные вебстраницы» (IN) -22 % ресурсов;
  - «конечные веб-страницы» (OUT), также 22 % ресурсов;
  - «отростки, мысы и перешейки» (22 % вебстраниц);
  - «острова», которые вообще не пересекаются с остальными ресурсами Интернет.

Неизменность пропорций и алгоритмов

- Было обнаружено, что пропорции названных категорий в течение нескольких месяцев оставались неизменными, несмотря на значительное увеличение общего объема веб-ресурсов.
- Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств веб-пространства, подтверждая тем самым наблюдение о том, что ***свойства структуры всего веб-пространства Bow Tie также верны и для его отдельных подмножеств.***
- Таким образом, алгоритмы, использующие информацию о структуре веб-пространства,

Закономерности модели Bow Tie

- Оказалось, что распределение степеней узлов (входящих и исходящих гиперссылок) веб-пространства (исследовались сайты домена edu в количестве 325729) подчиняется степенному закону, т.е. вероятность того, что соответствующая степень вершины равна *i*, пропорциональна $1/i^k$ (для входящих ссылок *к* «2.1, а для исходящих k=2.45).
- Кроме того, оказалось, что сеть WWW является «тесным миром» со средней длиной кратчайшего пути, равной 6, и относительно большим значением коэффициента кластерности, приблизительно равным 0,15 (для классического случайного графа это значение составило бы 0,0002) .

Ограничения модели Bow Tie

- Модель Брёдера не учитывает особенностей динамической части веб-пространства, формируемой потоками новостных сообщений. Применение модели «галстука-бабочки» к динамической составляющей веб-пространства нельзя считать корректным по ряду причин:
  - динамика информационных потоков влияет на природу гиперссылок, на сообщения, например, в течение определенного времени их может вообще не существовать;
  - модель Брёдера слабо учитывает особенности «скрытого» Web;
  - в информационных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части веб-пространства;
  - модель Брёдера не включает такого понятия как смысловое дублирование информации;
  - за прошедшее время с момента создания модели Брёдера появились новые разновидности гиперсвязей в веб-пространстве, например, существуют гиперссылки, доступные для пользователей-людей, но недоступные для роботов поисковых систем (в частности, определяемые тегом < noindex >).

Web 2.0

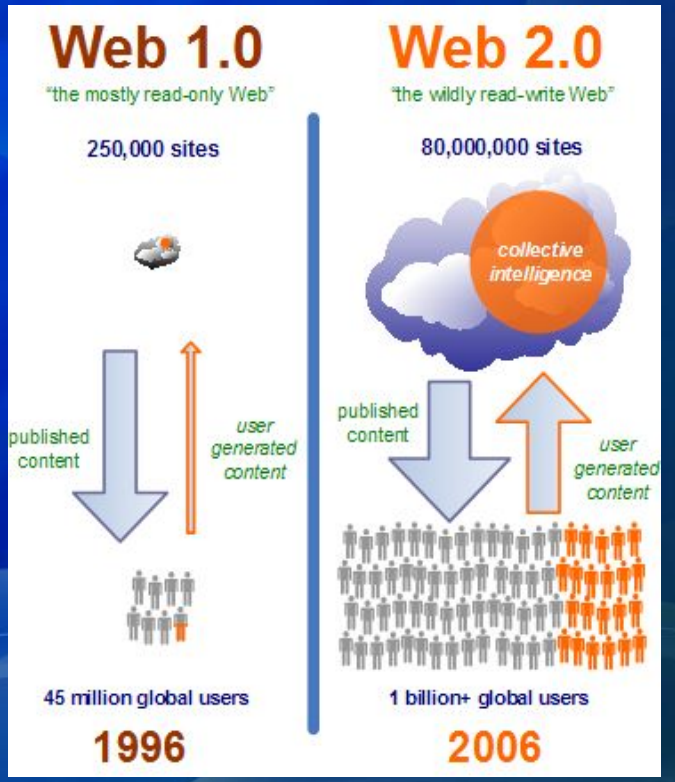
What is web 2.0?

- A concept made popular by Tim O'railey in 2004
- Basically – the web's move towards a "Prosumer" crowd
- Three main characteristics:

1) the growth of Web authoring styles that enabled many people to collectively create and maintain shared content;

2) the movement of people's personal on-line data (including e-mail, calendars, photos, and videos) from their own computers to services offered and hosted by large companies;

3) the growth of linking styles that emphasize on-line connections between people, not just between documents.



Different implications of web 2.0

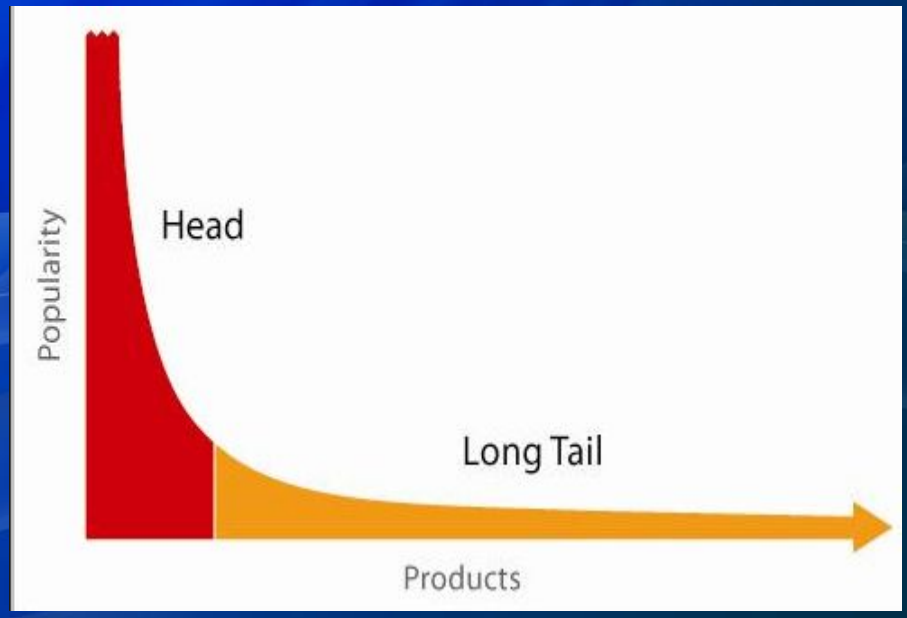
- Wikipedia grew rapidly during this period, as people embraced the idea of collectively editing articles to create an open encyclopedia on the Web (principle (1));
- Gmail and other on-line e-mail services encouraged individuals to let companies like Google host their archives of e-mail (principle (2));
- MySpace and Facebook achieved widespread adoption with a set of features that primarily emphasized the creation of on-line social networks (principle (3))

Different implications of web 2.0

- "Software that gets better as more people use it"
- "The wisdom of the crowds" - **Wisdom of Crowds** – By using this philosophy and encouraging user contributions in the form of feedbacks, reviews, rankings and user ratings.
- "The Long Tail" - business, marketing, internet.

The tail of a distribution represents a period in time when sales **for less common products** return a profit due to reduced marketing and distribution costs.

Long tail is when sales are made for goods not commonly sold. These goods can return a profit through reduced marketing and distribution costs.



Different implications of web 2.0

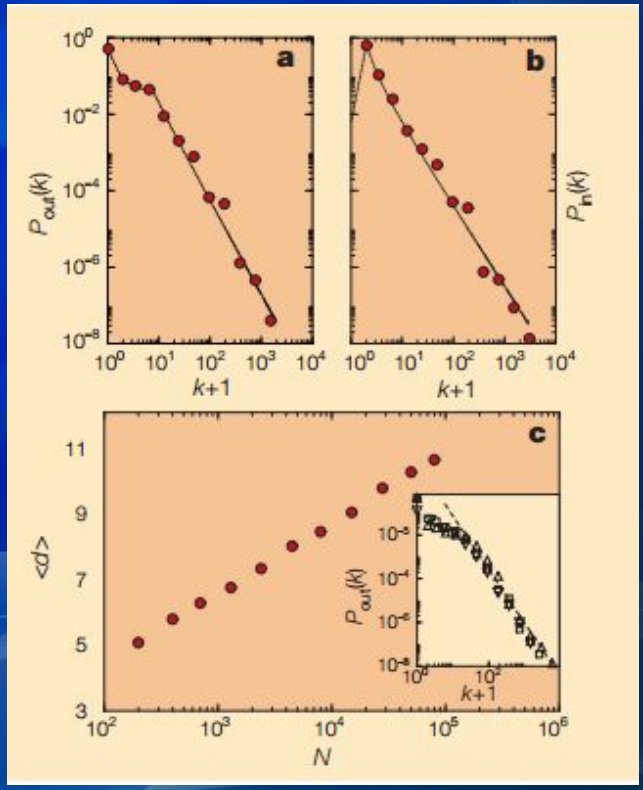
- The Long Tail («длинный хвост») — устоявшийся термин, пришедший из статистики и экономики. Впервые он был использован в октябре 2004 года Крисом Андерсоном (Chris Anderson) в статье журнала Wired. В статье отмечено: для многих новых экономик характерно существенное влияние продаж специфичных, нишевых продуктов, причем, прибыль от их реализации сопоставима с выручкой от продаж бестселлеров.
- Почти всегда в нишевых экономиках, достаточно подкрепленных спросом, можно построить успешный бизнес. Андерсон приводит несколько примеров. 57% от всех продаж книг интернет-магазина Amazon составляют «не-бестселлерные» книги, отсутствующие в большинстве «оффлайновых» книжных магазинов. 20% фильмов, взятых напрокат в Netflix на DVD, не идут на большом экране и не продаются в обычных магазинах. Более того, **суммарная стоимость малоизвестных товаров может оказаться на порядки выше стоимости «хитов»**.

A little bit more about the structure of the web

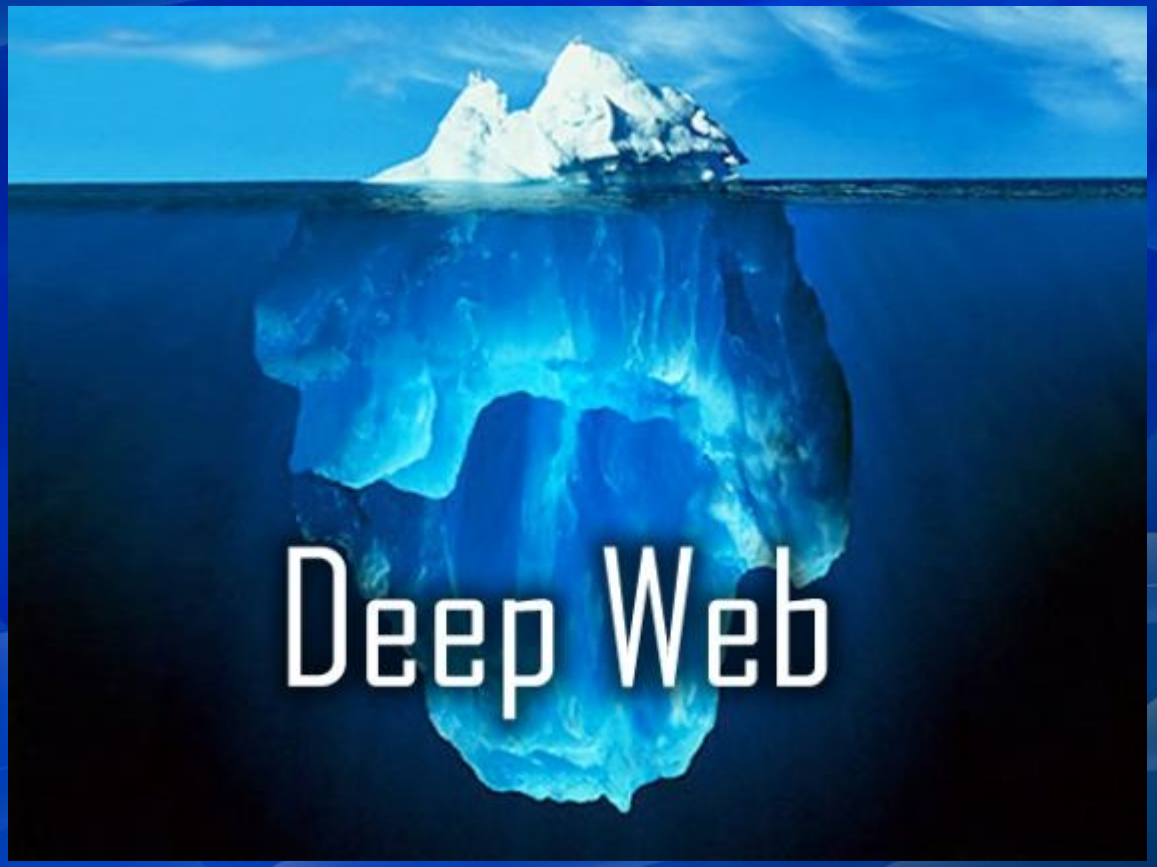
From: Albert R., Jeong H, & Barabasi A. - Diameter of the World Wide Web (2000)

About the research

- Trying to map reachability on the web
- Their main finding – the probability of a node to have k links (inbound and out) follow a **power law**
- Meaning – the web is a Small World Graph, typically found in biological and social networks
- This was proven more by the short path research



Невидимый WEB



Невидимый WEB

- Размер невидимого Интернета оценивается в 70% размеров сети.
- **Глубокая паутина** (также известна как **невидимая сеть**) — множество веб-страниц Всемирной паутины, не индексируемых поисковыми системами. Термин произошёл от *invisible web*.
- Наиболее значительной частью глубокой паутины является **глубинный веб** (*deep web, hidden web*), состоящий из веб-страниц, динамически генерируемых по запросам к онлайн базам данных.
- Не следует смешивать понятие *глубокая паутина* с понятием *тёмная паутина* (*dark web*), под которым имеются в виду сетевые сегменты, вообще не подключённые к сети Интернет.

Невидимый WEB

- С учетом изменений в вебе, которые произошли за последние десять лет, «невидимый» интернет грубо можно поделить на «персонифицированный интернет», «неиндексированный интернет» и «deep web».
  - «Персонифицированный интернет» — это интернет социальных сетей, типа Facebook, В Контакте и Google + с закрытыми для нефрендов страницами. При этом открытый контент Google+ индексируется соответственно Google, а Facebook – Bing`ом.
  - «Неиндексируемый интернет». Раньше значительную часть неиндексируемого интернета составляли страницы не html формата, т. е. файлы pdf, djvu, excel и т.п. К настоящему времени поисковики научились индексировать большинство указанных файлов и эта проблема отпала.
  - «Глубокий веб». Это значительная и очень интересная с точки зрения конкурентной разведки часть «невидимого интернета». К нему обычно относят сайты с динамическими страницами, требующими заполнения различного рода веб-форм, а также в ряде случаев, специальных паролей, логинов и т.п.

DEEP WEB

- Кроме полезного и содержательного с профессиональной точки зрения "Невидимого WEB" существует еще неиндексируемый интернет с невидимыми обычным поисковикам сайтами в основном криминального и антиобщественного характера. По факту, основная часть этого интернета принадлежит сети, развернутой на основе решения TOR.
- TOR был создан для своих нужд американской военно-морской разведкой. А позиционировался как "неподконтрольная никому альтернативная сеть."
- В настоящее время сеть TOR, несмотря на криминальный характер подавляющего числа сайтов, поддерживается несколькими крупнейшими некоммерческими фондами, а также рядом крупных американских корпораций и правительством Швеции.

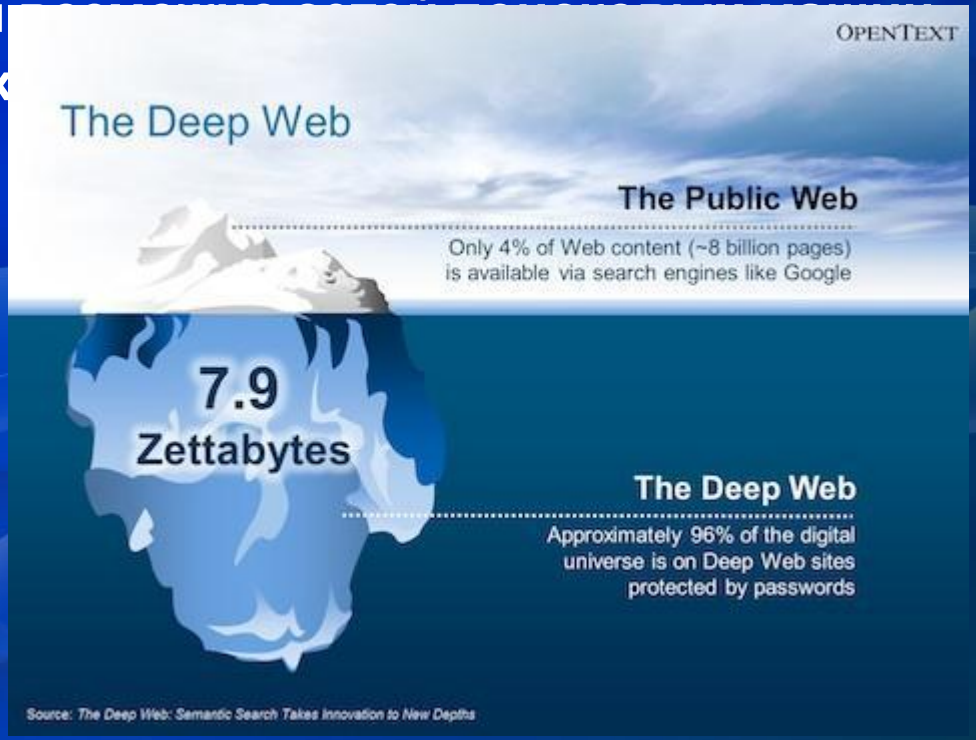
DEEP WEB

- Tor обеспечивает анонимное сетевое соединение, исключающее перехват данных и идентификацию пользователей посторонними. Анонимность трафика достигается за счет распределенной сети серверов.

- "Tor — это безопасное пространство для политических активистов, журналистов и других людей, за которыми потенциально могут следить. В основе Tor лежит принцип — не скрывать, а защищать. Ведь личные данные пользователя и то, что он делает в сети — это не секрет, это просто не для глаз посторонних." – одна из разработчиков - Руна Сандвик.
- Согласно отзывам – использовать Tor сложно.
- Это система прокси-серверов, позволяющая устанавливать анонимное сетевое соединение, защищённое от прослушивания. Рассматривается как анонимная сеть виртуальных туннелей, предоставляющая передачу данных в зашифрованном виде. Написана преимущественно на языках программирования C, C++ и Python.

Почему существует Невидимый Интернет?

- Поисковики не могут обнаружить несколько видов информации. Несколько наиболее актуальных примеров:
  - **Динамически создаваемые страницы**
  - **Исключенные страницы**
  - **Ограничения**
  - **Базы данных**



Почему существует Невидимый Интернет?

- **Динамически создаваемые страницы**
  - Это страницы, которые не могут быть получены на основе алгоритмов поисковых систем. Например, это касается поисковых запросов для сайтов. Которые содержат статистические данные. Поисковые алгоритмы не приспособлены для такого рода поиска и поэтому не могут дать пользователю доступ к таким данным. Примером является сайт Всемирного банка по статистике.

Почему существует Невидимый Интернет?

- **Исключенные страницы**
  - Некоторые владельцы сайтов предпочитают избегать появления тех или иных Web страниц сайтов в поисковых системах. Например, они не указывают метатеги, используют другие приемы. Это может быть проблемой, когда вы ищите конкретные технологии или компании, которые добровольно не хотят остаться «под радаром».

Почему существует Невидимый Интернет?

- **Физическое ограничение скорости.**
  - Поисковые машины имеют физические ограничения по скорости поиска новых страниц. Ежесекундно идет негласное соревнование: в Интернете появляются новые страницы, а поисковые машины наращивают свою мощь. Кроме добавления новых страниц, в Интернете происходят еще и исчезновение старых, а также внесение изменений в содержимое существующих, что также оттягивает на себя часть ресурсов поисковых машин. В этой постоянной гонке Интернет выигрывает у поисковых машин с большим перевесом.

Почему существует Невидимый Интернет?

Базы данных

- Большая часть мира структурированных данных была организована в БД, которые полностью доступны, но требуют от пользователя точно знать ключевые слова, чтобы найти информацию, в которой он нуждается. Примеры:
  - Статьи газет. Например «Financial Times», которые предлагают свободно для пользователей свои полные архивы.
  - European Patents, благодаря которому вы можете получить доступ к списку патентов, имеющихся в европейских патентных ведомствах.
  - Учредительные документы. При исследовании частных компаний обращение, например, к такому ресурсу, как Учредительные документы позволят вам получить доступ к информации о собственности, составе совета директоров и т.п. Подобные ресурсы имеются во многих государствах.
  - Финансовая информация. Вы можете найти ее, например, на

Почему существует Невидимый Интернет?

- Принцип попадания страниц в индекс при помощи пауков.
- Паук попадает только на те страницы, на которые есть ссылки с других страниц, либо которые внесены в очередь на индексирование вручную – путем заполнения формы «Добавить страницу» (“Add URL”). Соответственно, если на страницу никто не ссылался, и никто о ней не сообщал поисковой системе вручную, то такая страница не будет проиндексирована.
- Кроме того, если даже паук регулярно посещает страницу, то он делает это с определенной периодичностью. Если в промежутке между двумя посещениями страница изменится, то это изменение некоторое время будет неизвестно поисковой системе и ее пользователям. Таким образом, существуют две задержки по времени в индексировании страниц: когда страница создана, но еще неизвестна поисковой машине, и когда паук проиндексировал страницу, но не посетил ее

Инструменты и технологии работы в «Невидимом интернете»

- «Невидимый интернет» является наиболее интересной частью интернета не только для конкурентной разведки, но и для подавляющего большинства маркетологов, хэдхантеров, огромного отряда исследователей и ученых, то должны были появиться инструменты и технологии, которые позволяют работать в этой части Веба.
- Здесь будет приведено всего лишь несколько пример из огромного числа способов и инструментов поиска в deep web.
- Подавляющее число этих инструментов – англоязычные ресурсы и часто платные.

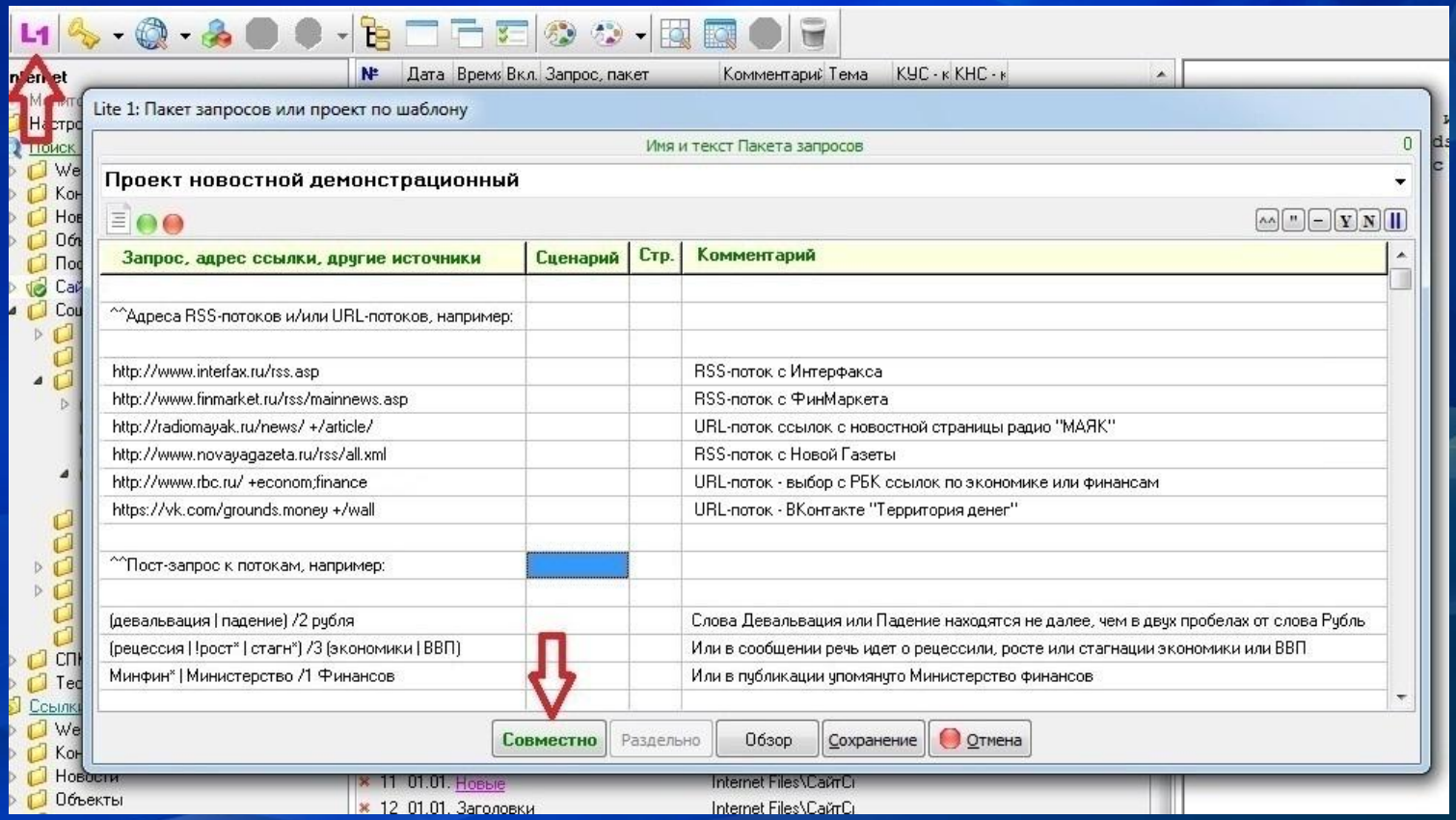
Инструменты и технологии работы в «Невидимом интернете»

К сожалению, нет нескольких русскоязычных инструментов (наши специалисты тоже не бездействуют), хотя и попытки эти носят все же локальный (личный), а не глобальный общественный характер.

- И первая ссылка на русскоязычный ресурс "Разведнет" - http://hrazvedka.ru/razvednet - содержит огромную рубрикацию инструментов глубинного поиска с комментариями.
- Хорошей востребованной программой для работы с неиндексированной частью «Невидимого интернета» является программа Алексея Мыльникова http://sitesputnik.ru, которая полностью позволяет сделать видимым неиндексированный интернет. Более того, эксперименты показывают, что дальнейшее развитие программы сможет решать вопросы придания видимости и бесплатной части «deep web».
- Такой же поиск могут осуществлять и специальные версии программы семейства Avalanche Андрея Масаловича.

Инструменты и технологии работы в «Невидимом интернете»

Программа Алексея Мыльникова SiteSputnik + Invisible. По ссылке http://sitesputnik.ru/ можно протестировать демоверсию программы.



Инструменты и технологии работы в «Невидимом интернете»

- В 2006 году Google получил патент на Поиск баз данных через формы-интерфейсы. Однако, как показали исследования Дмитрия Шестакова, применительно к сайтам Amazon.com и т.п. Google индексирует при помощи этого алгоритма не более 10% содержащихся в базе объектов. Повторенное недавно тестирование показало лишь незначительное увеличение до чуть более 15-17% этого показателя.
- В этих условиях некоторые компании, например, Brightplanet (http://www.brightplanet.com/) реализуют поиск в «deep web» как сервис.

Инструменты и технологии работы в «Невидимом интернете»

- Одновременно развивается целый ряд поисковиков, в основном связанных с текстовыми публикациями по самым различным отраслям бизнеса, науки и техники*). Фактически, это поисковые системы, сразу выходящие на конкретные базы данных и ведущие поиск в соответствии с заполненной веб-формой.
  - https://www.deepdyve.com/ - платный поисковик по глубокому вебу.
  - http://www.hozint.com/ - платформа для сбора информации о политической стабильности, безопасности стран, различных инцидентах и волнениях, собирающая и сканирующая информацию в глубоком вебе.
  - http://www.base-search.net/ – поисковик в невидимом вебе для открытых академических веб-ресурсов, принадлежащих лучшим университетам и исследовательским центрам США и Великобритании.
  - http://citeseer.ist.psu.edu/index – поисковик в невидимом вебе по различным публикациям, книгам, статьям в области компьютерных решений и информационных наук на английском языке.
  - http://www.deepwebtech.com/ – ведущий производитель «глубоких веб технологий», реализующих подход федеративного поиска, создатель

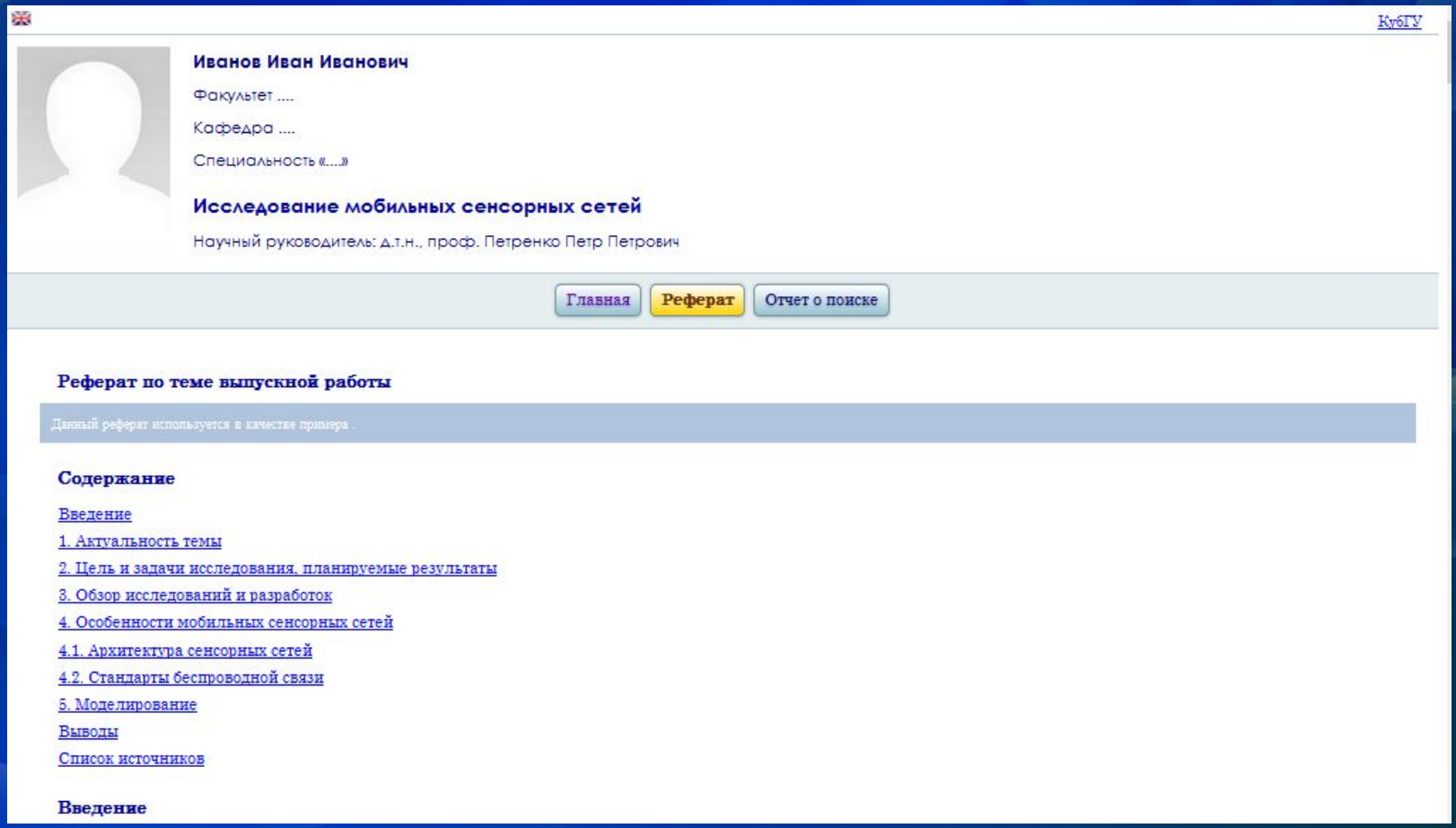
Инструменты и технологии работы в «Невидимом интернете» Задание № 1

- Выполнить поиск по тематике своей магистерской работы, используя:
  1. Демо-версию программы **FileForFiles & SiteSputnik (**sitesputnik.ru**)** Повторить поиск через несколько дней.
  2. Brightplanet (http://www.brightplanet.com/).
  3. Любой доступный инструмент из списка http://hrazvedka.ru/deepweb/deep-web.html
- Оформить результаты поиска в виде личной web-страницы, содержащей:
  1. Ваше имя, факультет, специальность, группу, имя руководителя, тему работы, цель работы,
  2. результаты тематического поиска, выводы о качестве инструментов глубокого поиска, сопоставимость результатов работы различных инструментов.
- Срок исполнения: 29 апреля – 6 мая.

Инструменты и технологии работы в «Невидимом интернете»

Задание № 1

- Образец личной страницы:



Literature

- Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze /An Introduction to informational retrieval. 2009 Cambridge UP. http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
- **Маннинг** К., **Рагхаван** П., **Шютце** Х. **Введение в информационный поиск**. Пер. с англ. –М.:"И.Д. Вильямс", 2011г. – 528с.-ил. ISBN 978-5-8459-1623-5.