Introduction to computational advertising

Bogdan Cautis

Outline

- **From IR to IS**
- **Advertising on the Web**
  - The evolution of Web advertising
  - Terminology
  - History
- **Advertising setting and problems**
  - Display ads
  - Sponsored search
  - Content match
- **The economics of Web advertising**
- **Conclusion**

From information retrieval to information supply

A brief history of Web search

- Based mainly on Information Retrieval (IR)
  - IR as a discipline has been around for 50-60 years
- June 11, 1994 – Brian Pinkerton: WebCrawler
- Dec 15, 1995 – AltaVista (crawled at 2.5 M pages per day, had 30 M pages)
- 1995, Yahoo (Yet Another Hierarchical Officious Oracle)
- 1998 – Google (googol 10^{100}), 2004 – Google IPO

Today

- «Without search engines the Web would not exist »

  (Andrei Broder, VP Emerging Search Technologies at Yahoo)

- Both a technology artifact and a social environment

- The normal way of life, search interaction is accepted paradigm

- Useful? For instance, it makes aggregation possible
  - Unlimited “selection stores” possible: Amazon

IR basics

- Corpus: Fixed document collection
- Goal: Retrieve documents with information content that is relevant to user's information need
- Relevance
  - For each query Q and stored document D in a given corpus assume there exists relevance Score(Q, D)

- **The context is ignored**
- **The users are ignored**

User needs

- Informational
  - want to learn about something (~40% / 65%)
- Navigational
  - want to go to that page (~25% / 15%)
- Transactional
  - want to do something (web-mediated) (~35% / 20%)

    - access a service

    - download

    - shop

Search engines generations

- First generation -- use only “on page”, text data
  - word frequency, language
- Second generation -- use off-page, web-specific data
  - link (or connectivity) analysis
  - sophisticated mathematical methods
  - click-through data (What results people click on)
  - anchor-text (How people refer to this page)
- Third generation -- answer “the need behind the query”
  - focus on user need, rather than on query
  - semantic analysis -- what is this about?
  - integrates multiple sources of data, context
  - **help the user!**

Third generation search engines

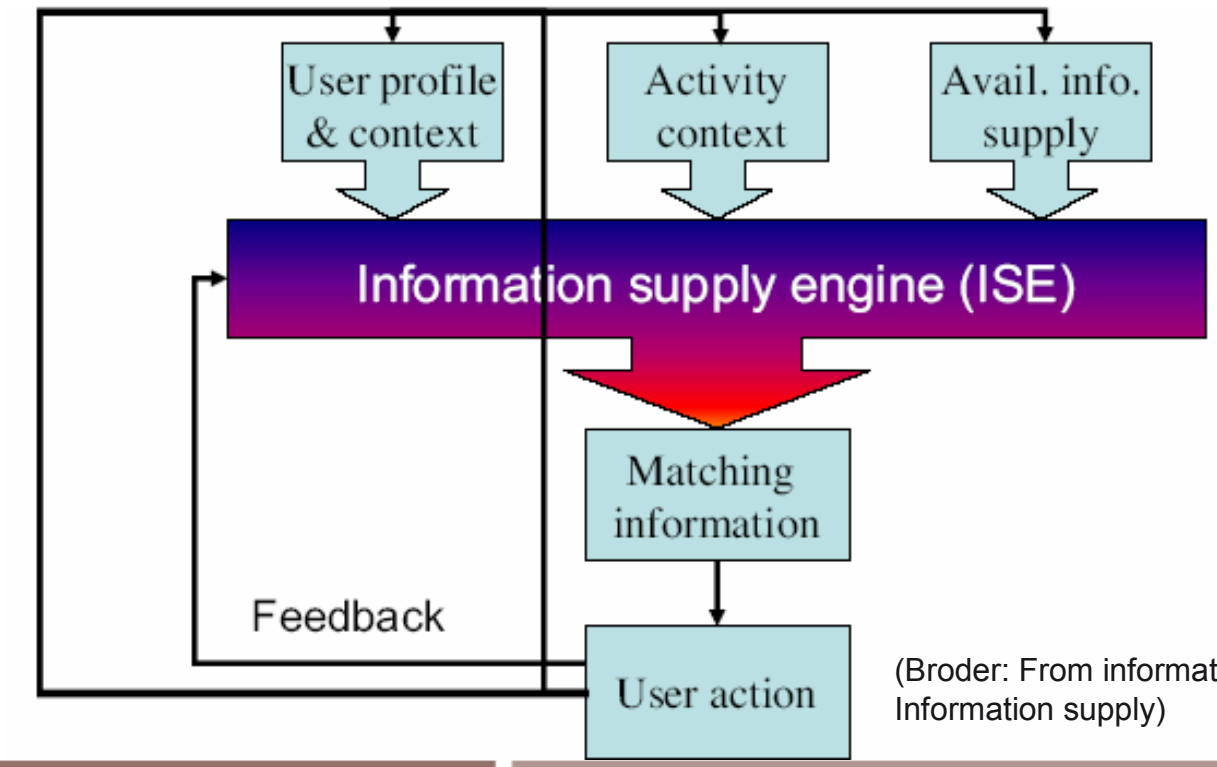
Understanding “the need behind the query” rather than simply returning query matches is

Triple win

- A win for users (better results)
- A win for content providers (focus)
- A win for search engines (monetization)

Information Supply

- **From information retrieval to information supply: Move from syntactic matching to semantic matching**



(Broder: From information retrieval to Information supply)

Examples

- Subscriptions (e-mail, RSS, etc)
- Alerts – News
- E-commerce sites: accessories, commentaries, related purchases, etc
- Automatic annotations
- Recommendations
- Contextual ads & search driven ads
- – …

Web advertising

Introduction

- **The Internet advertising spending is estimated over 17 billion dollars in 2006 ($150B total).**

- **97% of search revenue**
  - broadband is cheap, ubiquitous
  - “getting things done” easier on the Internet

- **Why does it work**
  - massive scale, automated
  - key: monetize more and better, “learn from the data”
  - new discipline: “Computational advertising”

Computational advertising

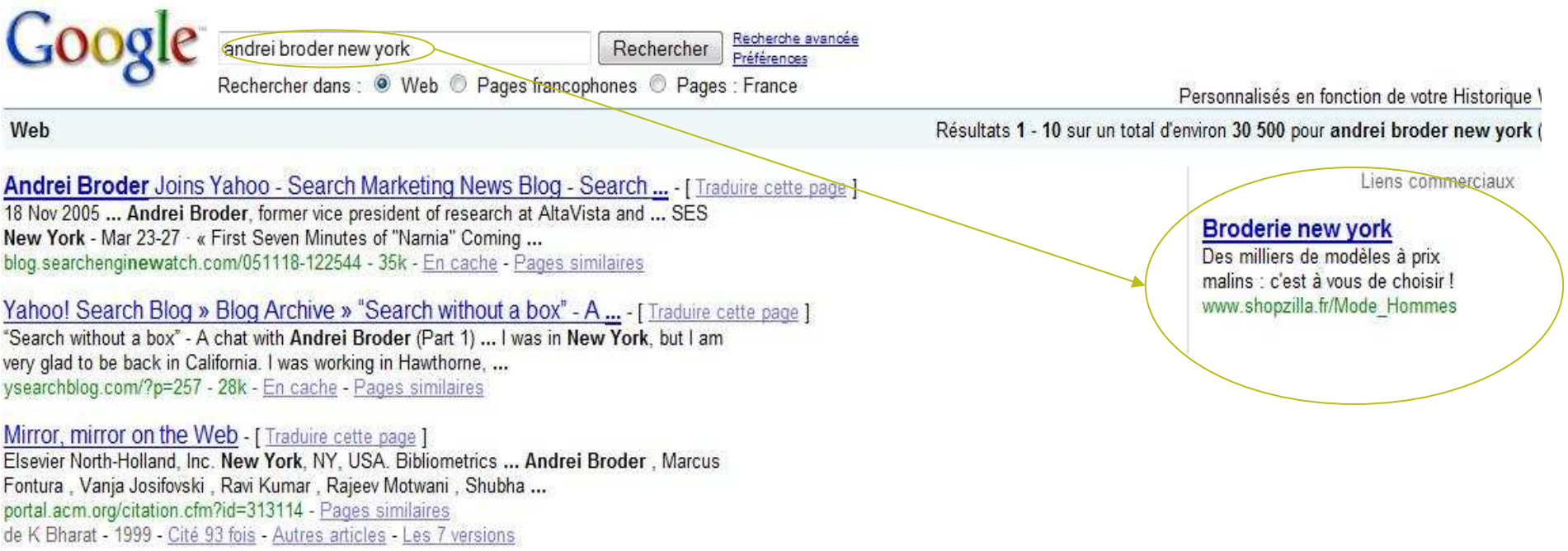
- **New scientific sub-discipline, at the intersection of**
  - large scale search and text analysis
  - information retrieval
  - statistical modeling
  - machine learning
  - optimization
  - microeconomics

Ad types

- **Three main types of textual Web advertising:**
  - ***Sponsored search*** which serves ads in response to search queries
  - ***Content match*** which places ads on third-party pages
  - ***Display advertising*** (banner ads)

- **Ads are information!**

A sponsored search ad



Another example

Google

theory | Rechercher | Recherche avancée | Préférences

Rechercher dans : Web Pages francophones Pages : France

Web Résultats **1** - **10** sur un total d'environ **211 000 000** pour **theory**

Lien commercial

Theory
LeBazarParisien.com/**Theory** Nouvelle Collection Printemps Eté Toute la Mode en Exclusivité !

Theory - [Traduire cette page]
IN **THEORY**, EVERYTHING IS POSSIBLE. **Theory** is a company, a concept, an aesthetic, a style and a product. **Theory** is a phenomenon that spun out of great **...**
www.**theory**.com/ - 6k - En cache - Pages similaires

Theory - Wikipedia, the free encyclopedia - [Traduire cette page]
16 Mar 2009 **...** For a more detailed account of theories as expressed in formal language as they are studied in mathematical logic see **Theory** (mathematical **...**

Liens commerciaux

Theory france
Vêtements mode et design ici.
Des offres hallucinantes à saisir!
www.ciao.fr/**Theory**

Theory
Des marques dans le vent
4000 articles dégriffés jusqu'à 70%
www.dress-for-less.fr

A content match ad (1)

Content match ad

(Broder: From information retrieval to Information supply)

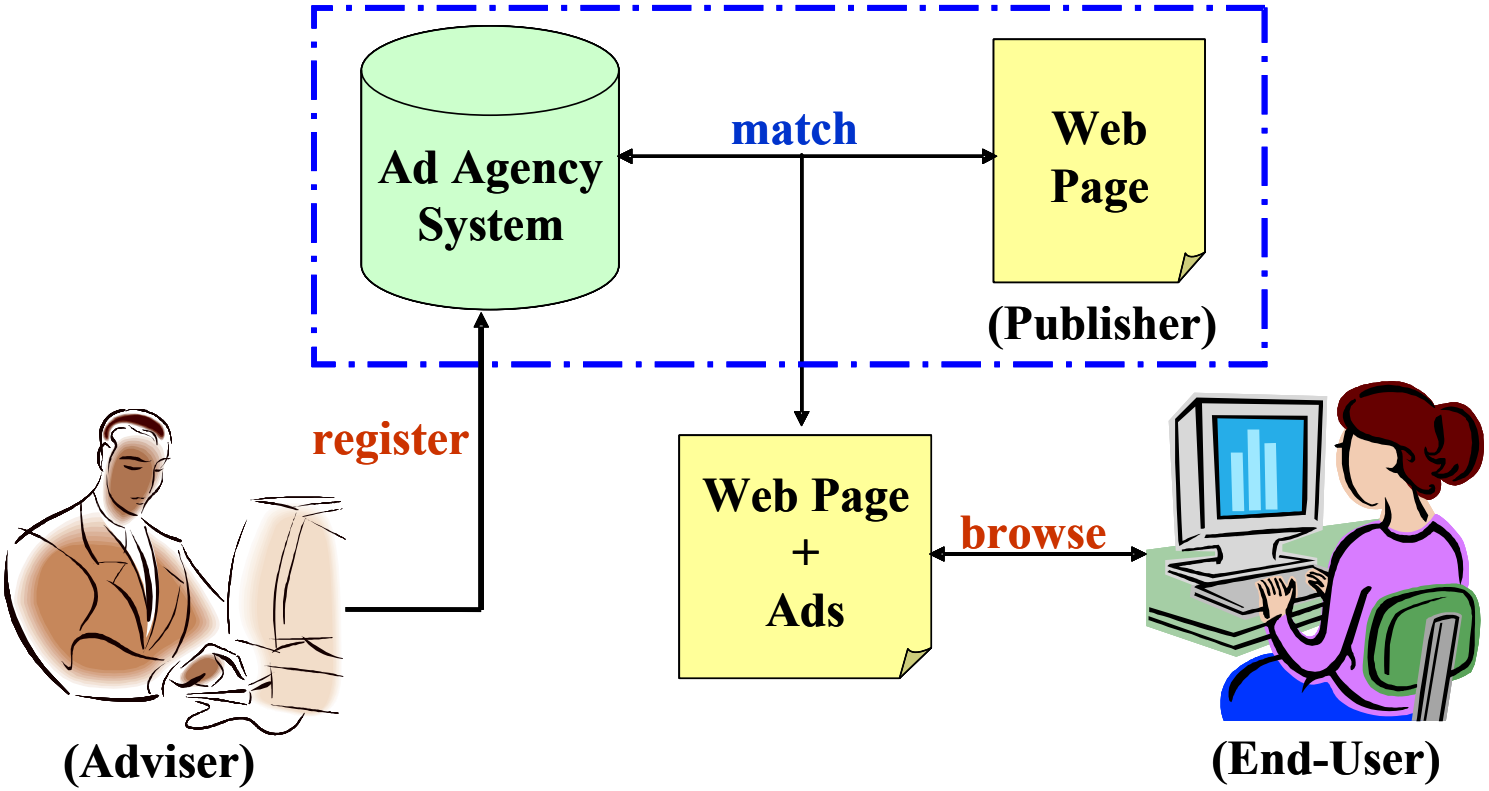
A content match ad (2)

(Broder: From information retrieval to Information supply)

Contextual Advertising Basic

- **Four interactive entities:**
  - The **publisher** is the owner of Web pages on which advertising is displayed.
  - The **advertiser** provides the supply of ads.
  - The **ad network** is a mediator between the advertiser and the publisher, who selects the ads that are put on the pages.
  - **End-users** visit the Web pages of the publisher and interact with the ads.

Overview of Ad display



Behind the curtains (sponsored search)

- Manual or automated review process to ensure that advertiser content is in fact relevant to the target keyword
- Matching advertiser content to user queries as they are received
- Displaying advertiser content in some rank order
- Gathering data, mesuring clicks, charging advertisers based on consumer clicks, etc.

Different revenue flavors

Revenue models

- CPM: Cost Per iMpression
- CPC: Cost Per Click
- CPV: Cost Per Visitor
- CPA: Cost Per Activity

Revenue models

- Example: suppose we show an ad N times on the same spot
- Under CPM: Revenue = N * CPM
- Under CPC: Revenue = N * CTR * CPC

Depends on the auction mechanism

Click-through rate

(probability of a click given an impression)

Revenue models

- Example: suppose we show an ad N times on the same spot
- Under CPM: Revenue = N * CPM
- Under CPC: Revenue = N * CTR * CPC
- Under CPA: Revenue = N * CTR * Conv. Rate * CPA

Conversion rate:

(probability of a user conversion on the advertiser's landing page given a a click)

Revenue models

Revenue dependence:

- **CPM: website traffic**
- **CPC: + ad relevance**
- **CPA: + landing page quality**

- **From 1st to 3rd: more relevant for advertisers, bigger prices and bids!**

Web advertising history

How it all started

- **In 1998, GoTo, later Overture Systems**
  - sponsored search
  - Yahoo acquired Overture in 2003
- **BeFirst followed in 1999**
- **Google adopted the model**
  - added click feedback (2002)
- **MSN extended the model to behavioral targeting**
- **HotWired introduced banner advertising in 1994**
- **The search engine OpenText offered preferred listings, in 1996.**

How it all started (2)

- First, CPM based (a.k.a. cost per mille)
- In 1996, Yahoo agreeed to charge Procter&Gamble by CPC
- By 1997, DoubleClick was offering CPA pricing
- In 2003, Overture, Google and FindWhat introduced automated tools for measuring CPA
- Auction models: in 1997, FlyCast and Narrowline
- Google's (2002) generalized second-price auction (GSP) in AdWords
- Pay the bid of the next highest bidder
- Later adopted by Yahoo!/Overture and others

Advertising settings and problems

Main issues

- What do you show to a user?
- How does the user interact with the ad system?

Display ads

- **Just pick ads**
  - graphically displayed
  - mostly for brandness awareness
  - revenue model is typically CPM

Display ads

- **Traditional advertising model:**
  - Ads are targeted at particular demographics
    - GM adds on Yahoo autos shown to « males above 55 »
    - Mortgage ad shown to « everybody Yahoo Front page »
  - Book a slot well in advance
    - « 2M impressions in Jan next year »
    - impressions guaranteed by the ad network!

Display ads

- **Fundamental problem: guarantee impressions to advertisers**
  - predict supply
    - how many impressions available
    - demographics overlap
  - predict demand
    - how much will advertisers want each demographic
  - find the « optimal » allocation
- **Forecast accuracy is critical!**

Content match and sponsored search ads

Content match

- Pick adds by matching them to content
- The user intent is unclear
- Webpage can be big and noisy

Sponsored search

- Given a search query
- Pick adds by matching them to the query
- User declares her intention
- Query is short and less noisy than Content Match

The main issues

- **Given a « query »**
  - Select the top-k ads to be shown on the k slots in order to maximize total expected revenue

- **What affects the total revenue**
  - Relevance of the ad to the query
  - Bids on the ads
  - User experience on the landing page (ad « quality »)

Implementation solutions

The data base approach (original Overture approach)

- Ads are records in a database
- The bid phrase (BP) is an **attribute**
- On query q
  - For exact match consider all ads with BP=q
  - For broad match rewrite q into “equivalent” queries q1, q2, … and consider all ads with BP=q1, BP = q2, …

The IR approach

- Ads are documents in an **ad corpus**
- The bid phrase is a meta-data
- On query q run q against the ad corpus
  - Have a suitable ranking function (more later)
  - BP = q (exact match) has high weight
  - No distinction between exact and broad match

Ad relevance computation

- **IR based**
  - use a search engine to match ads to context
    - ads are the « documents »
    - Context (user query or webpage content) are the query
  - Problem: word matches might not always work
  - Need to extract topical information
- **Machine learning from clicks**
  - Estimate CTR=Pr(click | ad, query, user)
  - Ad-ad similarity & collaborative filtering

Ranking idea

- Given a set of ads A_i
- Each A_i has a maximal bid B_i and an observed CTR C_i

Order ads by B_i*C_i decreasing

Usually, A_i pays less than B_i, just enough to beat the ad under it (second price auction)

Selection approaches

- **Exact match**
  - the ad's bid phrase matches the query
  - need query normalization
  - cannot bid on all feasible queries
- **Broad match: translate the query into bid phrases**
  - the ad platform finds good ads for a given query

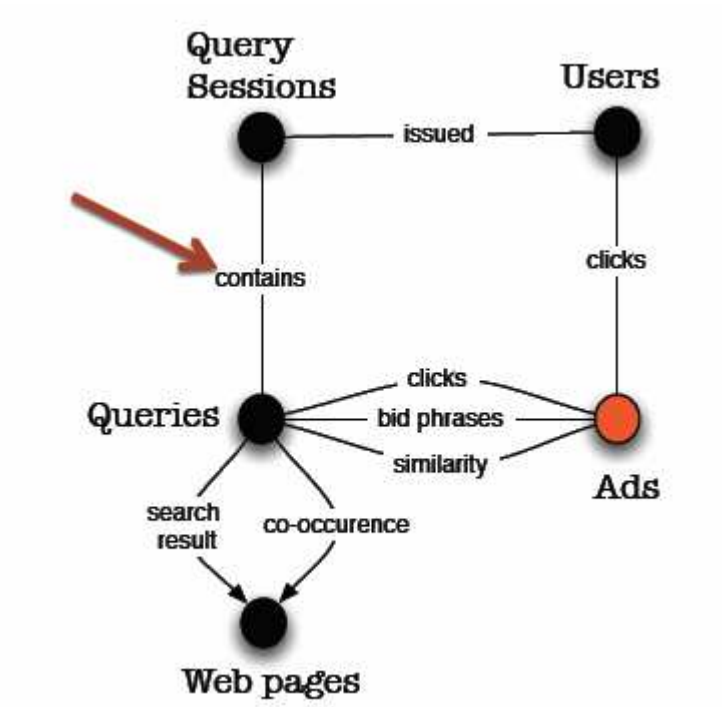
  (the advertiser did not bid on that specific keyword, but the query is deemed of interest to the advertiser)
  - pricing can be misleading
  - significant portion of the traffic has no bids …

- **Implementation**
  - Database lookup
  - Similarity search

Query rewriting

- Rewrite the user query q into Q' = (q_1, q_2, …)
- Use exact match to select ads for Q'
- **Offline vs online**
- Offline can be done only for queries that repeat often
  - More resources can be used
- Online
  - For rare queries offline not practical or simply does not work
  - Lot less time to do analysis (a few ms)
  - Limited amount of data (memory bound, time bound)

Rewriting using Web search logs



Query reformulations in a user session:

- insertions: game codes -> video game codes
- substitutions: john wayne bust -> john wayne statue
- deletions: skateboarding pics -> skateboarding
- spell correction: real eastate -> real estate
- specialization: jobs -> marine employment

Method: determine if p(rw|q) >> p(rw)

Rewriting using clicks

- Given a bipartite graph G:
  - V_q – nodes representing queries
  - V_a – nodes representing ads

- Edges connect queries with ads.
- Each edge has one or more weights
- For each pair of queries determine the similarity

$V = V_q \cup V_a$

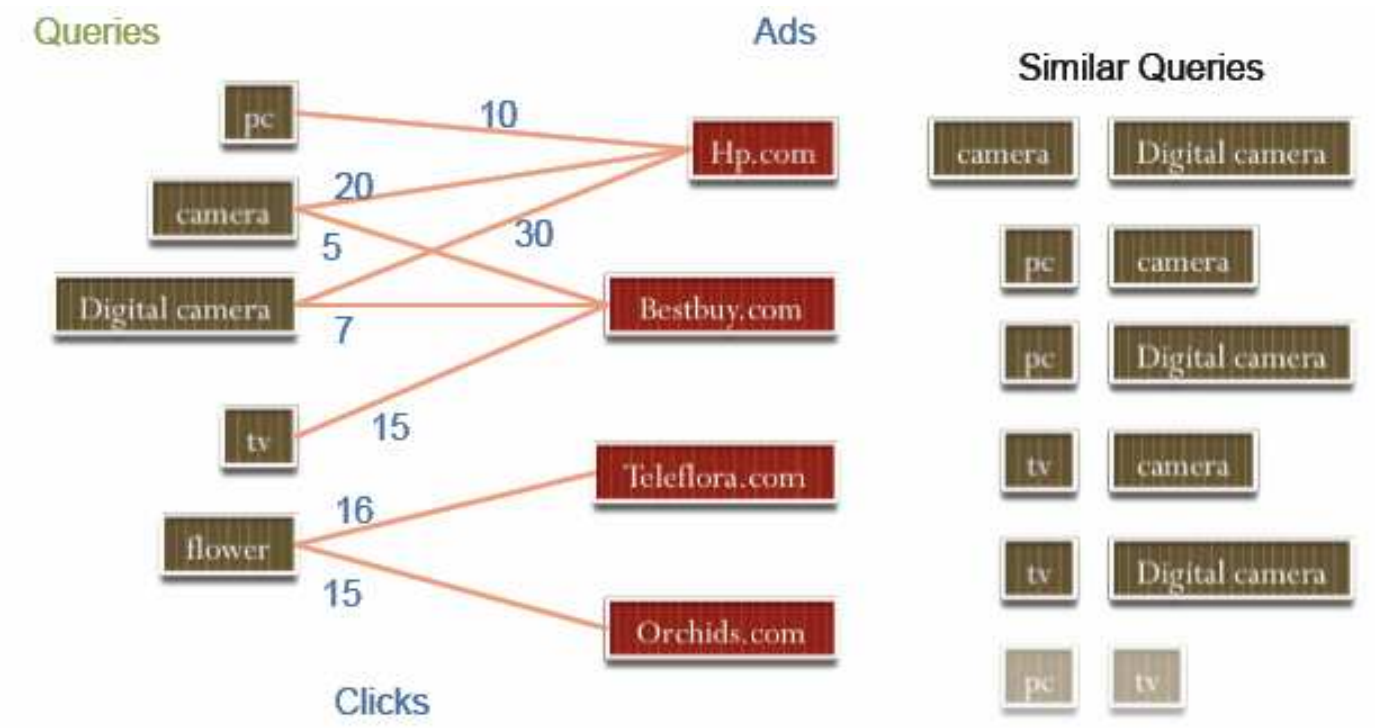
$E = \{e_1 \ldots e_k\}$ $e_i = (q,a,w)$ $q \in V_q$ $a \in V_a$ *w is a real number*

- Given pair of queries ($q1$,$q2$) find similarity $sim(q1,q2)$

On weights

- **Un-weighted**: there is an edge for each ad query pair where there is at least on click
  - some ads get a lot more clicks than others for the same query
- **Clicks**: weight the edges with the number of clicks on the (q,a) combination
  - Pairs with higher number of impressions get more clicks even if the relationship is not as strong
- **CTR**: keep the ratio between the clicks and impressions
  - CTR of 0.5 differs in confidence when we have 1 or 10K impressions

Example


Queries
Ads
Similar Queries
pc
camera
Digital camera
tv
flower
Hp.com
Bestbuy.com
Teleflora.com
Orchids.com
10
20
5
30
7
15
16
15
Clicks
camera
Digital camera
pc
camera
pc
Digital camera
tv
camera
tv
Digital camera
pc
tv

The Simrank algorithm (1)

- Intuition:

"Two queries are similar if they are connected to similar ads"

"Two ads are similar if they are connected to similar queries"

  - Assume similarity is a measure between 1 and 0 (like probability)
  - A query is "very" similar to itself: sim(q,q) = 1
  - Initially, we know nothing about the similarity with other queries:
  - sim(q,q') = 0 iff q ≠ q'
  - Establish similarity of two queries based on the ads they connect to
  - Then the same on the ad side
- Iterative procedure: at each iteration similarity propagates through the the graph

Simrank algorithm (2)

- E(q): set of ads connected to q
- N(q): # of ads connected to q
- simk(q,q'): q-q' similarity at iteration k

Start: sim(q,q) = 1, sim(q,q') = 0, sim(a,a) = 1, sim(a,a') = 0

$$sim_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} sim_{k-1}(i,j)$$

$$sim_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} sim_{k-1}(i,j)$$

- C – constant between 0 and 1, ensures diminishing impact with increased number of steps (small k sim goes to 0)

INSTITUT TELECOM

Example: first iteration

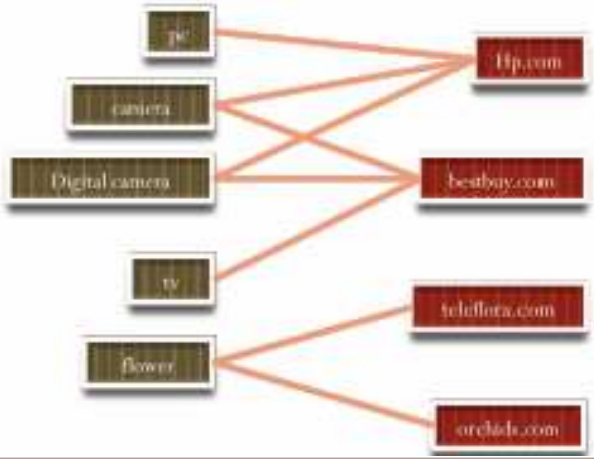
1st Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.0889	1			
digital camera	0.0889	0.1778	1		
tv	0	0.0889	0.0889	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



Example: second iteration

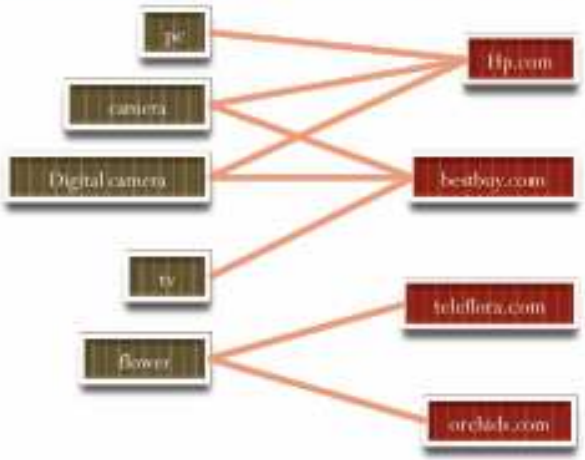
2nd Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.1244	1			
digital camera	0.1244	0.2489	1		
tv	0.0356	0.1244	0.1244	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



INSTITUT TELECOM

Example: 12th iteration …

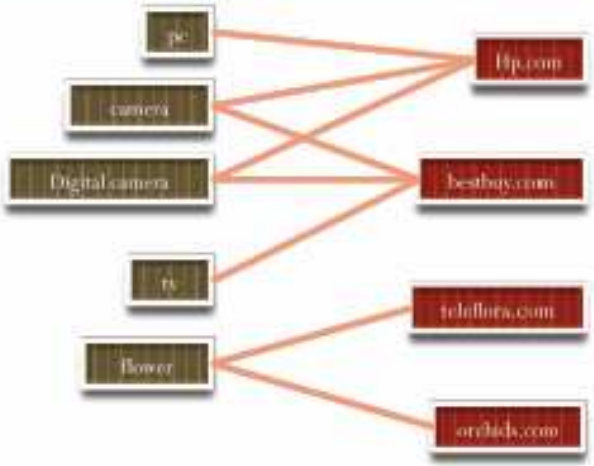
12th Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.1650	1			
digital camera	0.1650	0.33	1		
tv	0.0761	0.1650	0.1650	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



INSTITUT TELECOM

Big picture for sponsored search (IR view)

Ads corpus = Bid phrases + Title + URL + landing page + …

Ad query = Search keywords + context (location, user profile, search history)

Ad search is similar to web search but differences

- Ad database is smaller
- Ad database entries are small
- Ranking depends also on bids and CTRs

Ad relavance by online learning

- **Online learning**
  - previous approaches learned from historical data
  - slow response to emerging patterns,
  - initial biases never corrected

  if the system never showed “golf classes” for “iPod” it can never learn if this matching is good.

- **Solution: combine exploitation with exploration**
  - pick ads that are good according to current model
  - pick ads that increase your knowledge about the entire space of ads.

Online content matching

- **Web advertising for two types of Web page:**
  - ***Static page* (Offline)**: the matching of ads can be based on prior analysis of their entire content.
  - ***Dynamic page* (Online)**: ads need to be matched to the page *while it is being served to the end-user*. Thus, limiting the amount of time allotted for its content analysis.

Online content matching

- When a user views a page, the ad selection engine has only a couple hundred milliseconds to provide the ads.
- Offline approach: works well for static content pages that are displayed repeatedly.
- Ads need to be matched to the page while it is being served to the end-user, with extremely limited time for content analysis.

Big picture for content matching

Ads corpus = Bid phrases + Title + URL + landing page + …

Ad query = page + context (location, user profile, search history)

Similar to web search but differences

- Ad database is smaller
- Ad database entries are small
- Ranking depends also on bids and revenue
- The query is (current page) can be much larger than the target document

Collaborative filtering connection

Traditional IR based on fixed query-result correspondence

Ads: CTR probability replaces top-k results ->

- continuous CTR feedback for each (query, ad) pair
- learn the « best match between a user in a given context and a suitable advertisement »

- **data is sparse, in order to get the best match, we need to find similar ads, pages, and users.**

Dyadic interaction systems

- **Recommendation systems (user-movie, user-book)**
- **Web advertising (webpage/query-ad)**

A dyad is a pair (i,j):

- i=user, webpage, etc
- j=movies, ads

Measure some response: ratings, click-rates

- Other data: demographics, genres, etc

Goal: predict response to unknown dyads

The economics of Web advertising

Athens week, March 2010

INSTITUT TELECOM

Auctions

- Generalized First-Price Auctions 1997 auction revolution by Overture (then GoTo)
- Pay per-click for a particular keyword
- Links arranged in descending order of bids
- Pay your bid

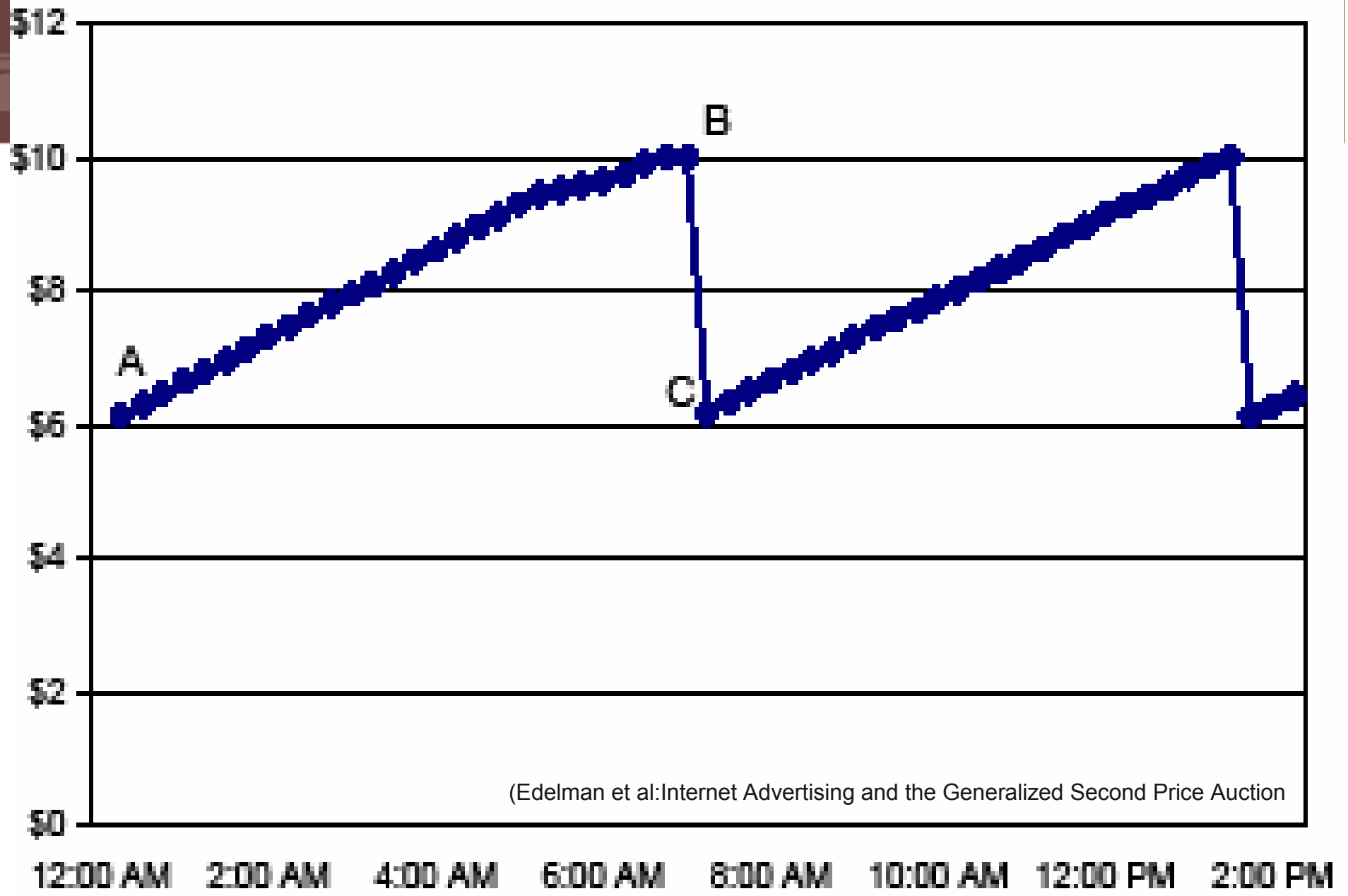
Problem: GFP is unstable because bids can be adjusted dynamically

Example on GFP

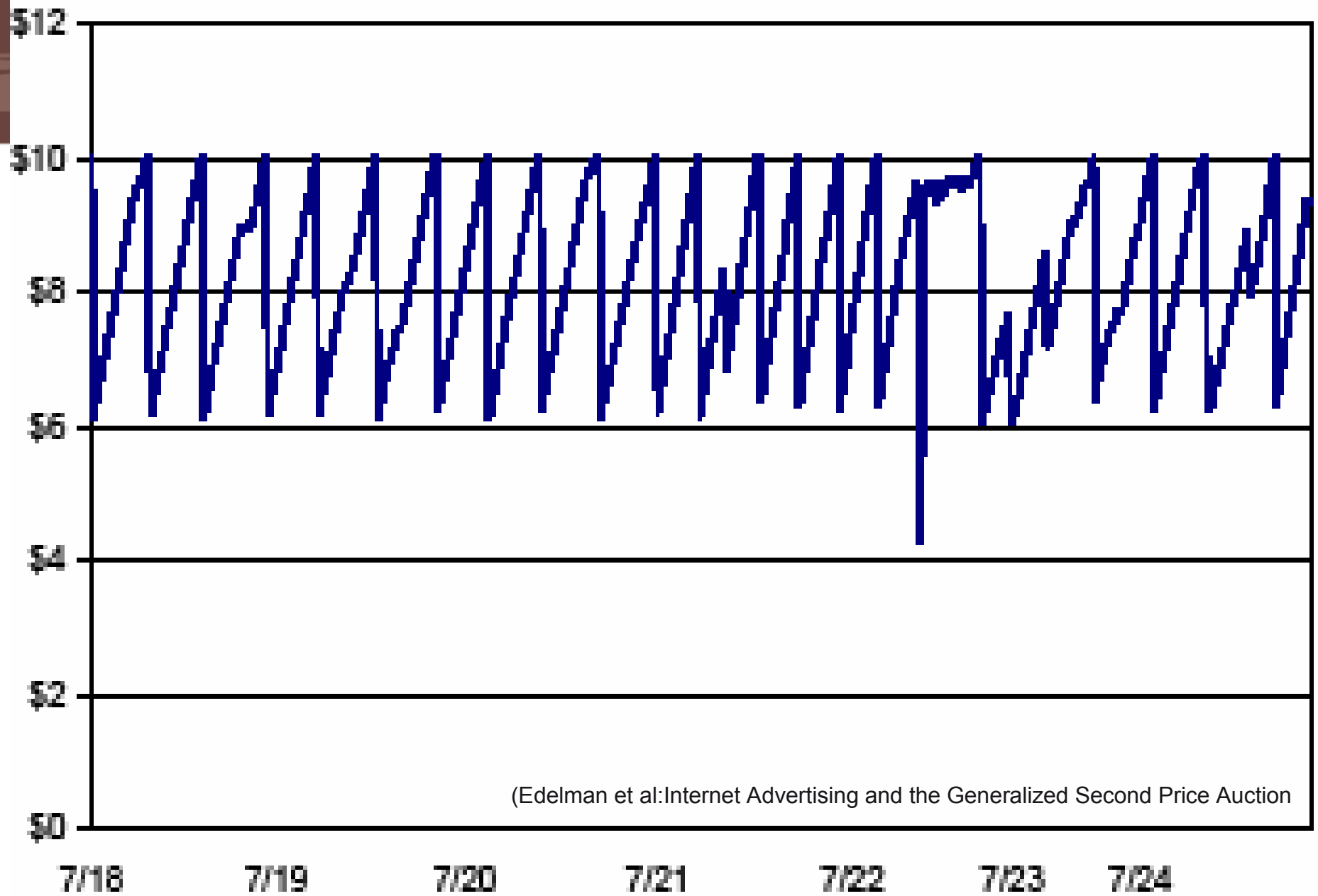
Two slots and three bidders.

- ad in first slot: 200 clicks per hour,
- ad in second slot: 100.

- bidders 1, 2, and 3 have values per: $10, $4, and $2
- if bidder 2 bids $2.01, to make sure he gets a slot.
- bidder 1 will not want to bid more than $2.02
- bidder 1 gets the top spot, but then bidder 2 will want to revise his bid to $2.03 to get the top spot,
- bidder 1 will in turn raise his bid to $2.04, and so on.



(Edelman et al:Internet Advertising and the Generalized Second Price Auction



(Edelman et al:Internet Advertising and the Generalized Second Price Auction

Generalized second-price auctions

- **A bidder in position i will never want to pay more than one bid increment above the bid of the advertiser in position (i + 1),**

- **Adopt this principle**
  - An advertiser in position i pays a price per click equal to the bid of an advertiser in position (i+1) plus a minimum increment (typically $0.01).

Conclusion

Key messages

- Computational advertising is a new scientific sub discipline that addresses the problem of finding the best match between a given user within a context, and a set of advertisements
- Is a $20 billion+ industry, still early in the technical and business model maturity cycle
- Two main types of online advertising are graphical and textual advertising
- Textual ads are information items and as such IR techniques can be used for ad selection
- Sponsored search is the main channel for textual advertising on the web
- Ads are selected in sponsored search using an exact match to the bid phrase or advanced match to the whole ad
- Main ad selection approaches are the database approach (lookup for exact match) and the IR approach where we look up using multiple features
- Query rewrite is a common advanced match technique where the query is rewritten into another query that is used to retrieve the ads by exact

Many active research areas & open problems

- query understanding
- content matching
- sentiment analysis
- online modeling
- massive optimization
- text summarization
- named entity extraction
- computer-human interaction
- economics of ads

References

- Computational Advertising course @ Stanford:
  http://www.stanford.edu/class/msande239/
- Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords, *Edelman, Ostrovsky and Schwartz*
- From query based Information Retrieval to context driven Information Supply, *Andrei Broder*
- Algorithmic Challenges in Online Advertising, *Deepak Agrawal and Deepayan Chackrabarti*
- Just in time contextual advertising, *Anagnostopoulos et al.*
- Internet Advertising and Optimal Auction Design, *Schwarz*