CS 277, Data Mining

Web Data Analysis: Part 2, Advertising

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Internet Advertising, Bids, and Auctions

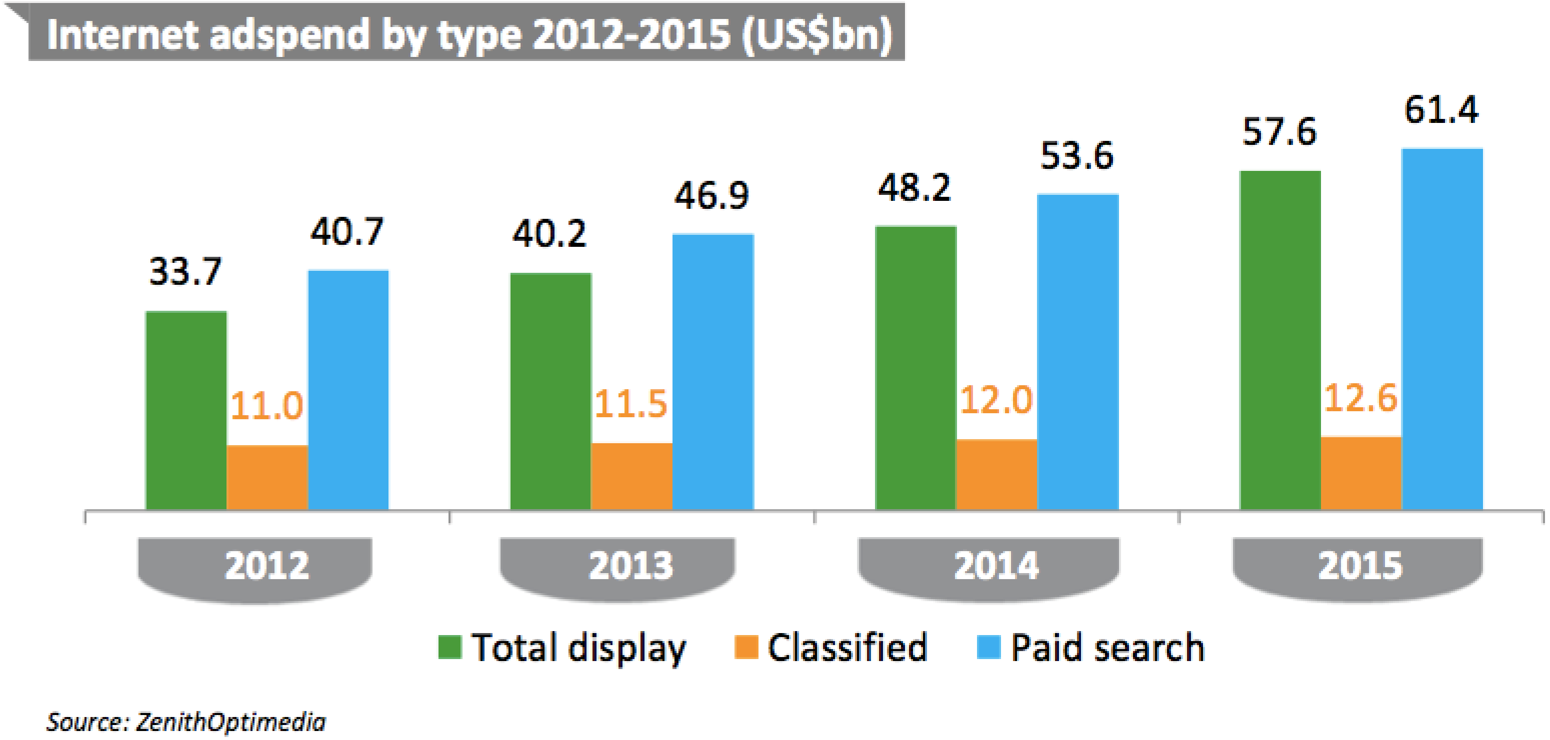
"Computational Advertising"

- Revenue of many internet companies is driven by advertising

- Key problem:
  - Given user data:
    - Pages browsed
    - Keywords used in search
    - Demographics
  - Determine the most relevant ads (in real-time)
  - About 50% of keyword searches can not be matched effectively to any ads
  - Other aspects include bidding/pricing of ads

- New research area of "computational advertising"
  - See link to Stanford class by Andrei Broder on class Web site

Why is Advertising Important for Internet Companies?



From Techcrunch.com, Sept 30, 2013

Types of Online Ads

- Display or Banner
  - Fixed content, usually visual
  - Or (more recently) video ads

- Sponsored search (Text Ad)
  - Triggered by search results
  - Ad selection based on search query terms, user features, click-through rates, ….

- Context-based/Text (Text Ad)
  - Can be based on content of Web page during browsing
  - Ad selection based on matching ad content with page content

Participants in Online Advertising

- Publishers
  - Provide the space on Web pages for the ads
  - e.g., Search engines, Yahoo front page, CNN, New York Times, WSJ

- Advertisers
  - Provide the ads
  - e.g., Walmart, Ford, Target, Toyota…

- Ad Exchanges
  - Match the advertisers and publishers in real-time
  - e.g., Doubleclick, Google, etc
  - Contract with advertisers to run advertising campaigns, e.g., deliver up to 100k clicks using up to 10 million impressions in 30 days
  - Ad-server runs complex prediction/optimization software (in real-time) to optimize revenue (from ad-server's viewpoint)

Concepts in Online Advertising

- Impression: showing an ad to an online user
  - CTR = clickthrough rate (typically around 0.1%)

- Revenue mechanisms (to ad-exchange or publisher, from advertiser)
  - CPM: cost per 1000 impressions
  - CPC: cost per click
  - CPA: cost per action (e.g., customer signs up, makes a purchase..)

- Ad-exchanges and auctions
  - Impressions can be bid on in real-time in ad-exchanges
  - Typically a 2nd-price (Vickery) auction
  - Key to success = accurate prediction of CTR for each impression

U.S. INTERNATIONAL 中文网

?

The New York Times

?

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather

WORLD U.S. NEW YORK BUSINESS OPINION SPORTS SCIENCE ARTS FASHION & STYLE VIDEO All Sections

?

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS 59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 6:47 PM ET
- VIDEO: **Confrontation in Crimea**

An Obama Budget Big on Ideals, but With Small Chances

By JACKIE CALMES 9:02 PM ET

Some Who Fled Cuba Are Returning to Help

By DAMIEN CAVE 8:55 PM ET

Some members of the first Cuban families to leave after Fidel Castro took over are coming back, reuniting with the island and

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- **Op-Docs: 'Chinese, on the Inside'**

MARKETS » At 10:03 PM ET

JAPAN Nikkei	HangSeng	CHINA Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

Get Quotes | My Portfolios »

?

Each ? represents an “ad slot”

In real-time the ad-exchange will compute which ads to show a particular user


U.S. INTERNATIONAL 中文网

SHOP MARCJACOBS.COM MENS WATCHES

The New York Times

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather

WORLD U.S. NEW YORK BUSINESS OPINION SPORTS SCIENCE ARTS FASHION & STYLE VIDEO All Sections

HERE'S TO A NEW YEAR RECEIVE 50% OFF The New York Times WINE CLUB BUY NOW>

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS 59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- Kerry Takes Offer of Aid to Ukraine 33 minutes ago
- Cyberattacks Rise as Crisis Spills to Internet 6:47 PM ET
- VIDEO: Confrontation in Crimea

An Obama Budget Big on Ideals, but With Small Chances

By JACKIE CALMES 9:02 PM ET

Some Who Fled Cuba Are Returning to Help

By DAMIEN CAVE 8:55 PM ET

Some members of the first Cuban families to leave after Fidel Castro took over are coming back, reuniting with the island and

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- Editorial: Frustration With Afghanistan
- Brooks: Putin Can't Stop
- Cohen: Russia's Crimean Crime

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- Op-Docs: 'Chinese, on the Inside'

MARKETS » At 10:03 PM ET

JAPAN Nikkei	HangSeng	CHINA Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

Get Quotes | My Portfolios »

INTRODUCING TODAY'S PAPER WEB APP

The newspaper experience in digital form

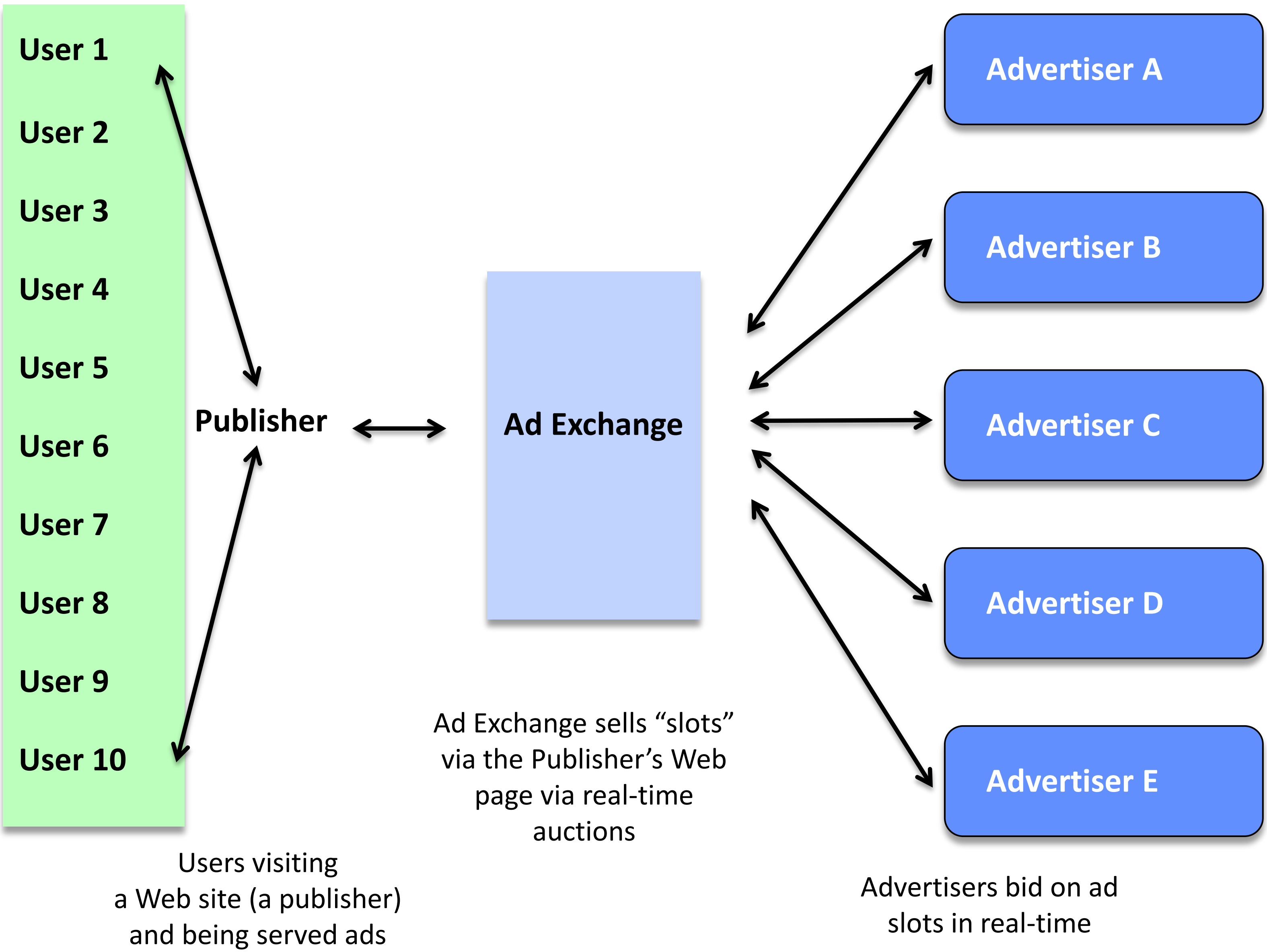
GO TO TODAY'S PAPER >

FREE TO DIGITAL AND HOME DELIVERY SUBSCRIBERS

The New York Times

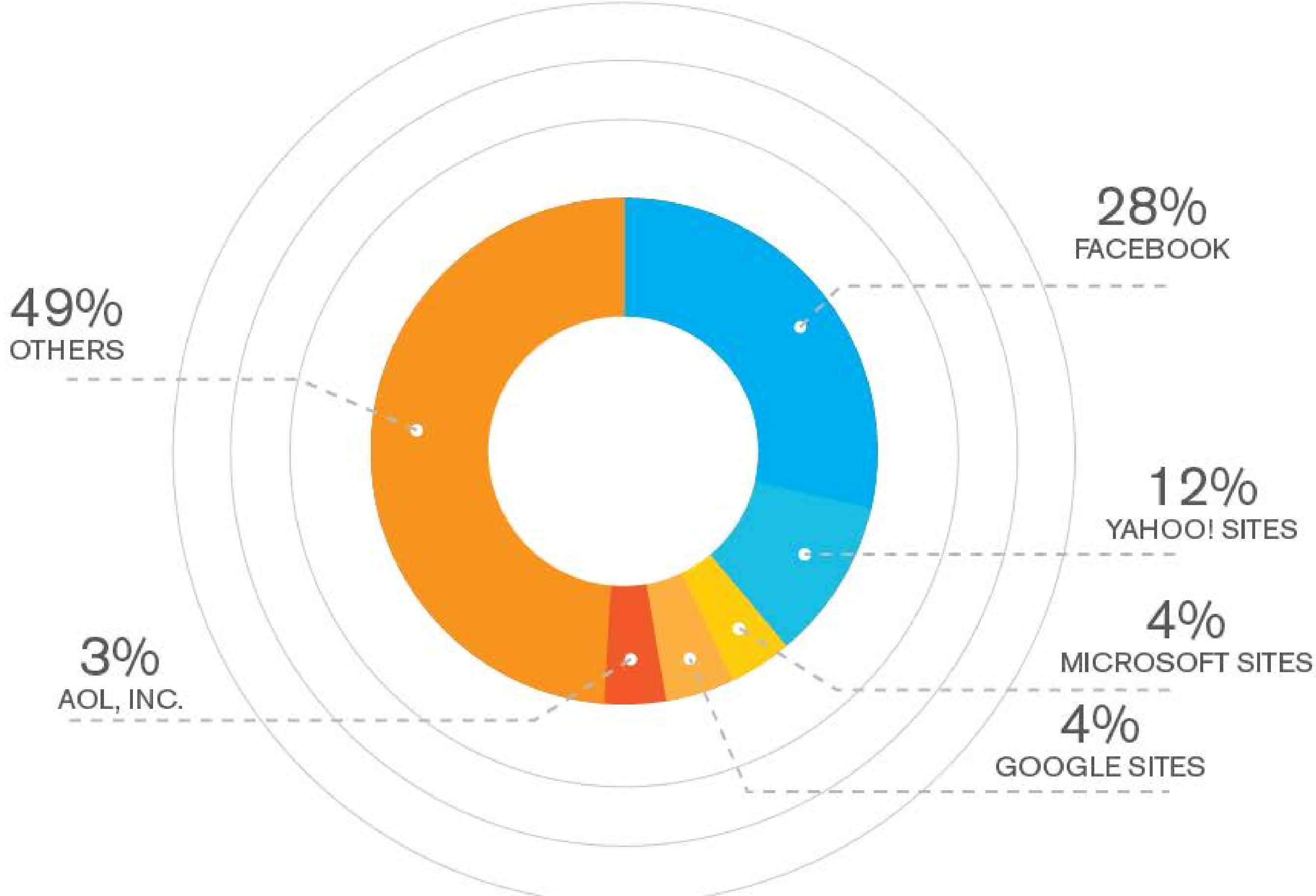
These ads are "impressions"

Simplified View of Advertising (Publisher View)



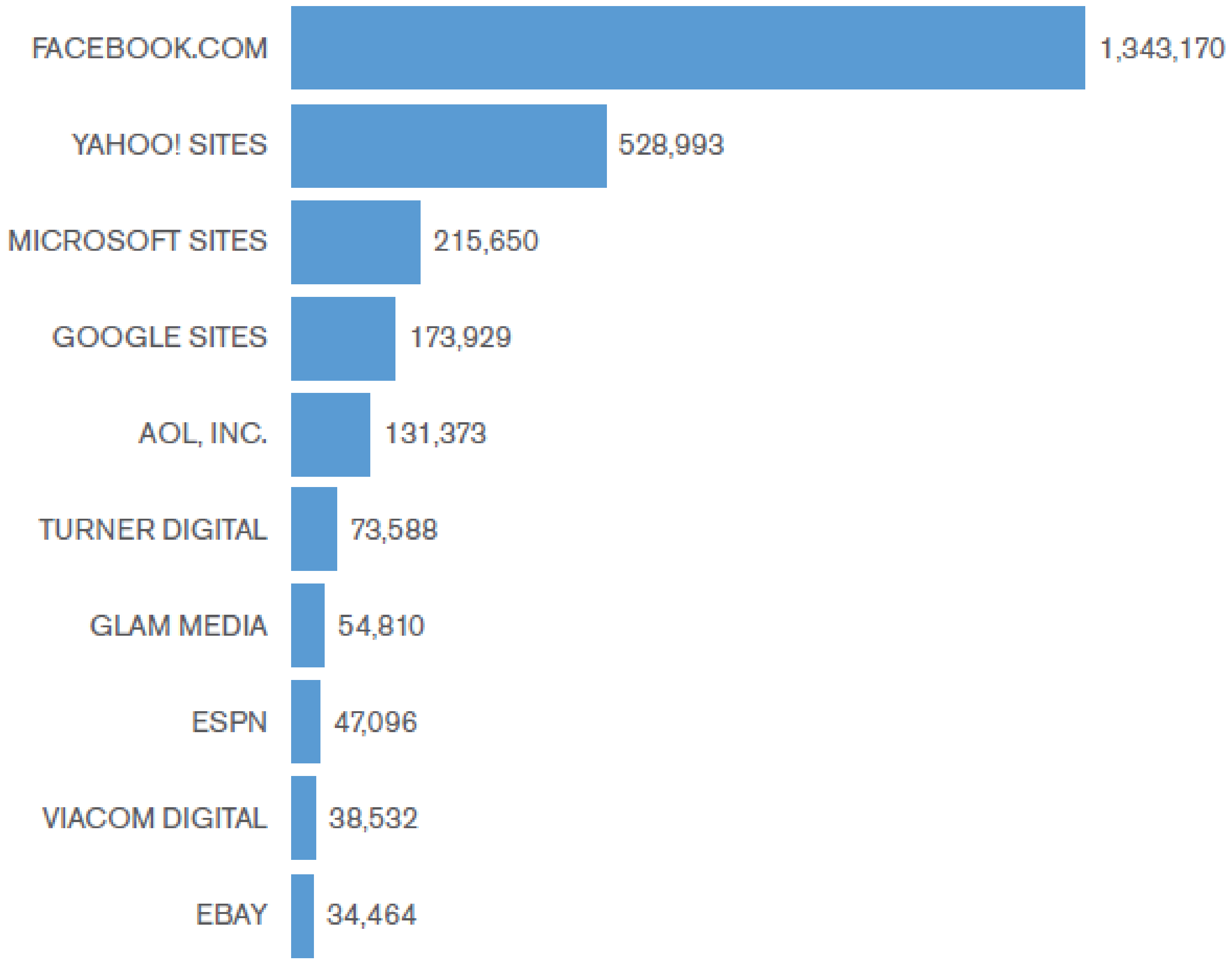
Publisher Share of Display Ad Impressions

Source: comScore Ad Metrix, U.S., Q3 2011

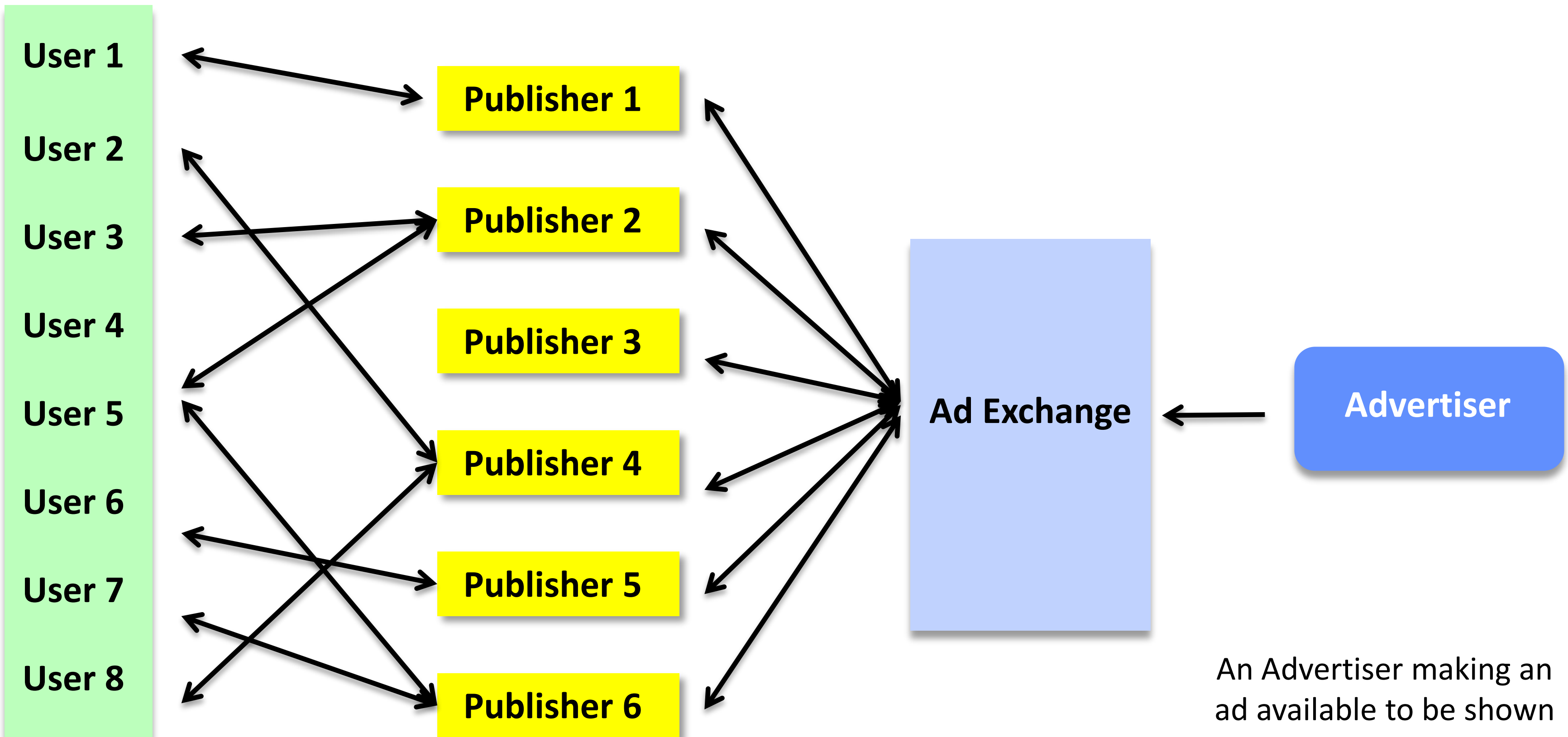


Top Ten U.S. Online Display Ad Publishers by Number of Impressions in Millions

Source: comScore Ad Metrix, Jan-2011 to Dec-2011, U.S.



Simplified View of Advertising (Advertiser View)



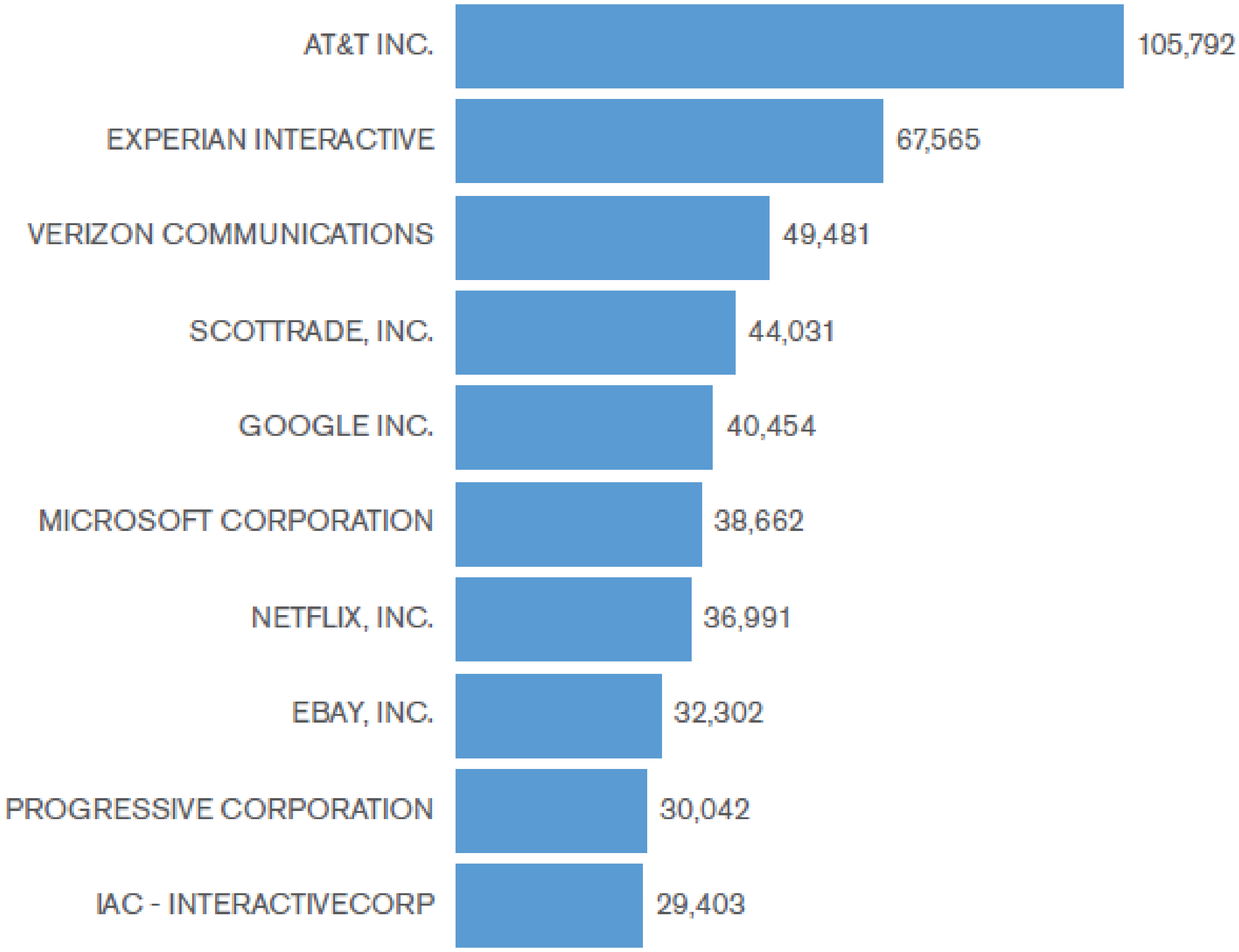
An Advertiser making an ad available to be shown to some set of users

Users visiting Web sites (publishers) and being served ads

Publishers selling "inventory" (ad slots) on an Ad Exchange

Top Ten U.S. Online Display Advertisers by Number of Impressions in Millions

Source: comScore Ad Metrix, Jan-2011 to Dec-2011, U.S.



UCIrvine
University of California, Irvine

Behind the Scenes…

- The previous slides are a very simplified picture of how these systems work……… in practice there are many other factors

- Multiple 3rd party "advertising companies"
  - In practice rather than just a single "ad exchange" there is a whole "ecosystem" of different systems and companies that sit between the publisher and the advertisers, optimizing different parts of the ad matching process

- Auction mechanisms
  - Use of "2nd price auctions"

Auctions and Bidding for Queries

- Say we have a query (like "flower delivery")
- Different advertisers can bid to have their ad shown whenever this search query is entered by a user
- Say there are K different positions on the search results page, each with different likelihood of being seen by user
  - For simplicity imagine that they are in a vertical column with K positions, top to bottom
- Advertisers submit bids (in real-time) in terms of how much they are willing to pay the search engine for a click on their ad (CPC model)
  - Tradeoff between the getting a good position and paying too much
- So there is an auction (often in real-time) among the advertisers

Auction Mechanisms

- Initial Internet advertisers paid flat fees to search engines (per impression)

- Overture (later purchased by Yahoo!) in 1997 introduced the notion of bidding and auctions
  - Advertisers submitted bids indicating what they would pay (CPC) for a keyword
  - Improvement over flat fees…..but found to be inefficient/volatile, with rapid price swings, which discouraged advertisers from participating

- 2002: Google introduced the idea of 2^{nd} price Auctions for keyword bidding
  - Advertisers make bids on K positions, bids are ranked in positions 1 through K
  - Advertiser in position k is charged
    the bid of advertiser in position k+1 plus some minimum (e.g,. 1 cent)
  - Advertiser in Kth position is charged a fixed minimum amount
  - Google (and others) quickly noticed that this made the auction market much more stable an “user-friendly”, much less susceptible to gaming
    (Yahoo!/Overture also switched to this method)
  - Google’s AdWords uses a modified ranking:
    - Instead of ranking by Bid it ranks by Bid * Estimated CTR

Example of 2nd Price Auction Bidding Work?

- 2 slots and 3 advertisers
  - So the advertisers want to (a) get a slot, and (b) get the best slot

- Advertisers place a true value on a click of $10, $4, $2 respectively
  - This notion of "true value" is important
  - It is what an advertiser truly believes a click on their ad is worth
  - Or in other words, it is the maximum they should be willing to pay

- 2nd price auction: each advertiser bids their true value
  - Advertiser 1 is ranked 1st, gets slot 1, and pays $4 + 1 cent
  - Advertiser 2 is ranked 2nd, gets slot 2, and pays $2 + 1 cent
  - Advertiser 3 is ranked 3rd and gets no slot

2^{nd} Price Auctions

- Various economic arguments as to why this is much more efficient than 1^{st} price auctions
  - Advertisers have no incentive to bid anything other than their true value
  - This discourages advertisers from dynamically changing bids, which was a cause of major instability in earlier first-price auctions

- Methods seems to work particularly well for internet advertising

- References:
  - Edelman, Ostrovsky, and Schwarz, American Economic Review, 2007
  - H. Varian, Online Advertising Markets, American Economic Review, 2010

Google's second price auction

Note that the rank here is based on Bid * CTR

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate: when an ad is displayed, what percentage of time do users click on it? CTR is a measure of relevance.
- ad rank: bid × CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- rank: rank in auction
- paid: second price auction price paid by advertiser

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent).

Slide from Heinrich Schutze, Introduction to Information Retrieval Class Slides, University of Munich, 2013

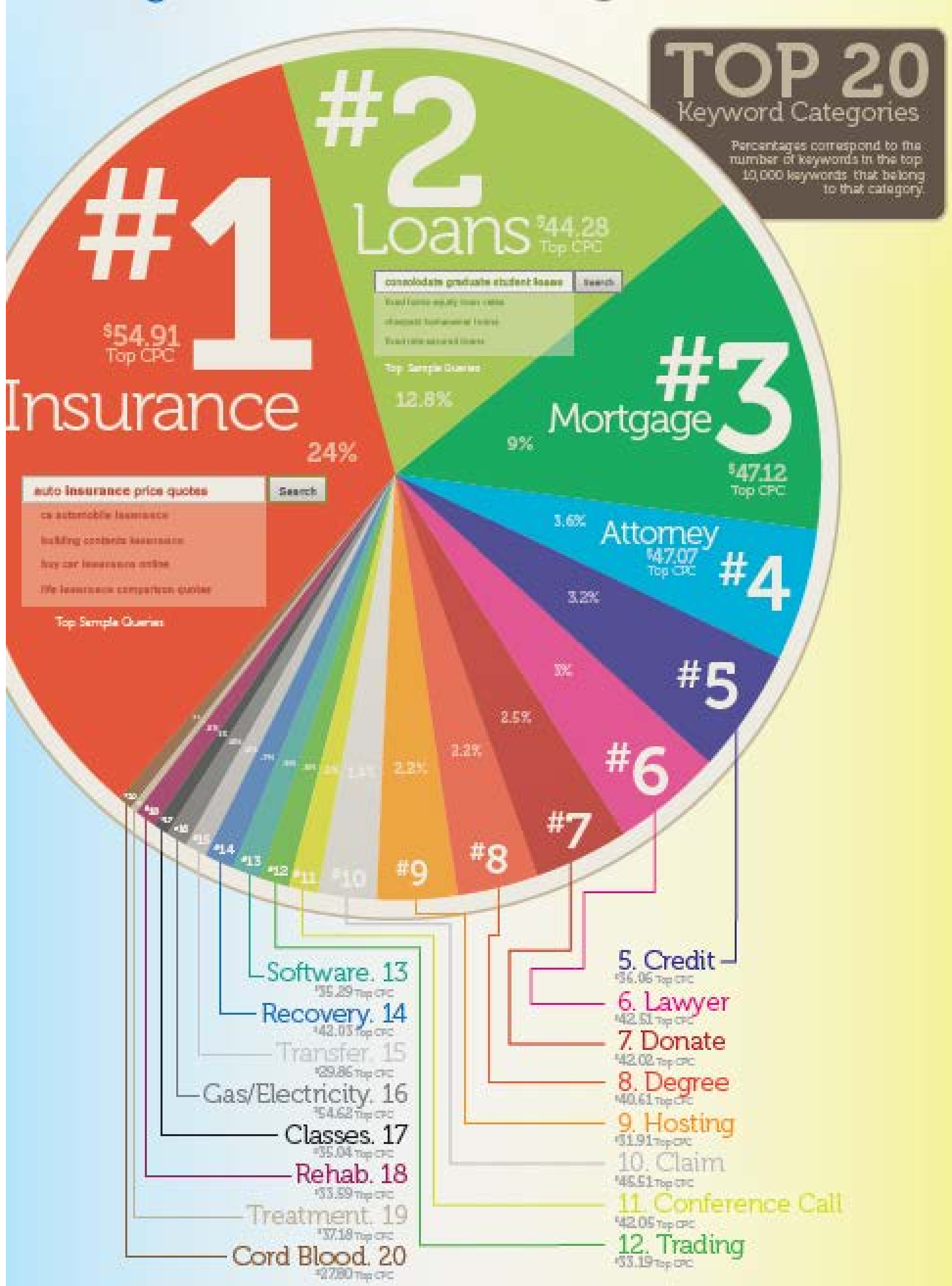
Keywords with high bids

According to http://www.cwire.org/highest-paying-search-terms/

\$69.1	mesothelioma treatment options
\$65.9	personal injury lawyer michigan
\$62.6	student loans consolidation
\$61.4	car accident attorney los angeles
\$59.4	online car insurance quotes
\$59.4	arizona dui lawyer
\$46.4	asbestos cancer
\$40.1	home equity line of credit
\$39.8	life insurance quotes
\$39.2	refinancing
\$38.7	equity line of credit
\$38.0	lasik eye surgery new york city
\$37.0	2nd mortgage
\$35.9	free car insurance quote

Slide from Heinrich Schutze, Introduction to Information Retrieval Class Slides, University of Munich, 2013

Top 20 most expensive keywords in Google AdWords Advertising



Source: http://www.wordstream.com/download/docs/most-expensive-keywords.pdf

Examples of Costs per Click

Metric	2010	2011	2012	2013
Cost per click (CPC)	$1.24	$1.04	$0.84	$0.92
Click through rate (CTR)	0.7%	0.4%	0.5%	0.5%
Average Ad Position	3.7	3.0	2.6	2.1
Conversion rate	6.8%	5.3%	3.4%	8.8%
Cost per conversion	$13.14	$19.74	$24.40	$10.44
Invalid click rate	6.7%	10.9%	8.0%	8.3%

From: survey data from 51 advertisers,
at http://www.hochmanconsultants.com/articles/je-hochman-benchmark.shtml

Predicting Click-Through Rates for Online Advertisements

Optimally Matching Advertisements to Users

- Advertising is a very large component of revenue for search engines
  - Displaying the "best" set of ads to users is a key issue

- Problem Statement (from search engine's perspective)
  - Inventory = a set of possible ads that could be shown
  - Query = query string typed in by a user
  - Problem: what is the best set of ads to show the user, and in what positions

- This is a complicated optimization problem
  - Objectives:
    - Search engine: maximize revenue (usually by attracting clicks)
    - Advertiser: maximize click rate
    - User: only wants to see relevant ads (overall user quality)
  - Other aspects
    - Each advertiser may only want to show a fixed maximum number of ads
    - User saturation if they see the same ad multiple times
    - Click fraud, etc

Cost-Per-Click (CPC) Model

- Cost-Per-Click, or CPC:
  - Search engine is paid every time an ad is clicked by a user

- Simple Expected Revenue Model

$$E[\text{ revenue }] = p(\text{click} \mid \text{ad}) \; CPC_{\text{ad}}$$

- Simple heuristic
  - Order the ads in terms of expected revenue

Examples of Costs per Click

Metric	2010	2011	2012	2013
Cost per click (CPC)	$1.24	$1.04	$0.84	$0.92
Click through rate (CTR)	0.7%	0.4%	0.5%	0.5%
Average Ad Position	3.7	3.0	2.6	2.1
Conversion rate	6.8%	5.3%	3.4%	8.8%
Cost per conversion	$13.14	$19.74	$24.40	$10.44
Invalid click rate	6.7%	10.9%	8.0%	8.3%

From: survey data from 51 advertisers,
at http://www.hochmanconsultants.com/articles/je-hochman-benchmark.shtml

Expected Revenue Model

- Simple Expected Revenue Model

$$E[\text{ revenue }] = CTR_{ad} \times CPC_{ad} = p(\text{click} \mid \text{ad})\ CPC_{ad}$$

- CPC_{ad} is known ahead of time: the key problem is estimating CTR

- Typically we also condition on additional factors beyond the ad itself, e.g.,
  - We really want to estimate p(click | ad, query, user, ad_position)
  - For simplicity we will ignore everything except "ad" here

- If we have some click data we can just estimate

$$P(\text{click} \mid \text{ad}) = (\text{number of clicks}) / (\text{number of times ad was shown})$$

- Typical click through rates are small, e.g., 1 in 1000 or 1 in 10000
  - So we are typically trying to estimate the probability of a rare event

Computing the CTR from Click Data

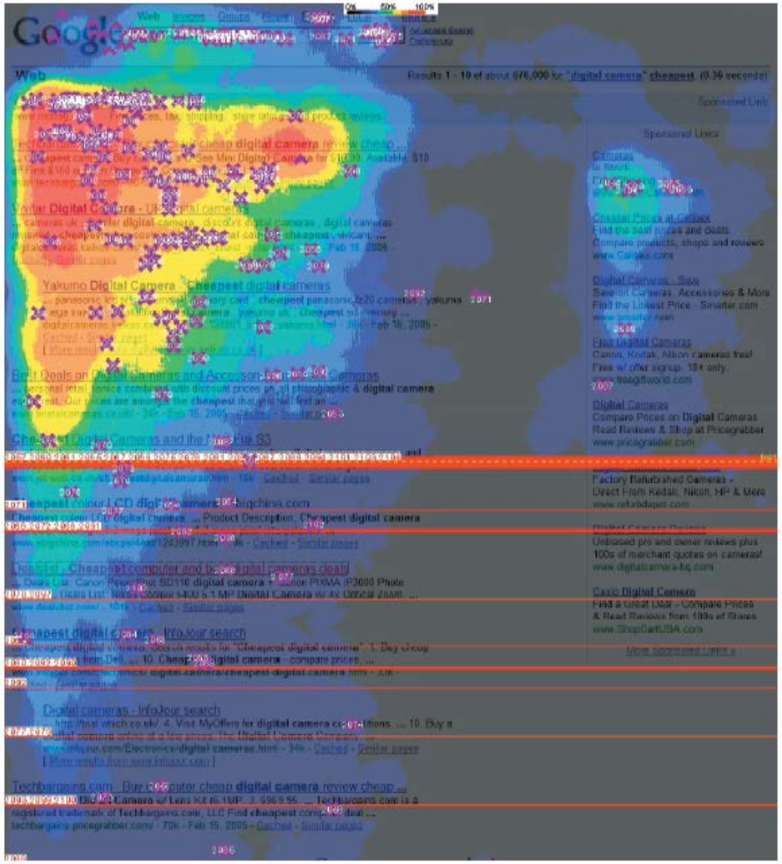
- Estimate of CTR = (number of clicks)/(number of views)

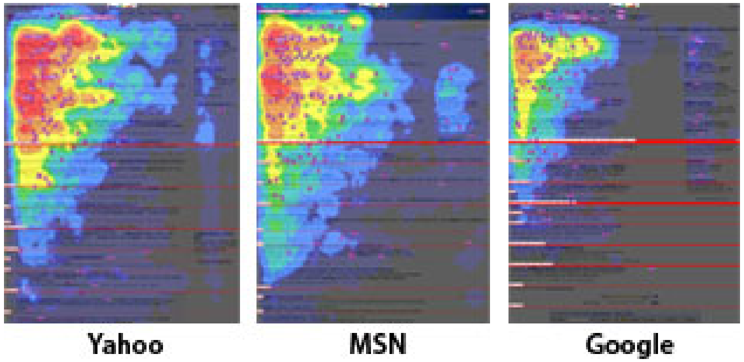
- Number of clicks = number of times ad was clicked

- Number of views?
  - Use a "discount" model based on eye-tracking to estimate how many times the ad was seen by users
  - So number of views is total number of times ad was shown, "discounted" by position model

Eye-Tracking: The Golden Triangle for Search

from Hotchkiss, Alston, Edwards, 2005; EnquiroResearch





Simple Example of CTR Estimation

- Assume that the true P(click | ad) = 10^{-4}
  - Say we have seen r clicks, from N showings of the ad
  - Our estimate of P(click | ad) = P' = r/N

- What is our uncertainty about P'?

Simple binomial model, assume N p > 5, i.e., N > 5 x 10^4 in our problem

-> 95% confidence interval is

$$w = 1.96\ \sqrt{p\,(1-p)}/\sqrt{N} \quad \approx \quad 0.02/\sqrt{N}$$

Say we want w < 10^{-5} (10% of the true value)

Rearranging terms above this means we need

$\sqrt{N}$ > 0.02 10^5 or N > 4 x 10^6

This means we need a very large N to be confident in our estimation of small probabilities

Difficulty of CTR Prediction Problem

- Clickthrough rates are small -> need large number of impressions to get reliable estimates
- Every day there will be a large number of new ads that the ad placement algorithm has not seen before, i.e., with unknown CTR
- Making mistakes is expensive
  - Say we show ad A 10 million times, and the CPC is $1 with a true CTR of 10^{-4}
  - And we don't show ad B, which has a CPC of $1 with a true CTR of 10^{-2}
  - Then the "cost of learning" about ad A (versus not showing B) is
    10^{-2} times 10 million, or $100,000 (!)

Online Learning of ClickThrough Rates

Online Learning of CTRs

- Once we begin to show ads, we would like to learn the CTRs

- Consider K different ads, with CTRs of p_1, ….. p_K

- We would like to learn these CTRs so that we can maximized expected revenue……but we don't want to lose too much potential revenue in doing so

- This is an example of the "explore/exploit" problem
  - Explore: for each ad show it enough times so that we can learn its CTR
  - Exploit: once we find a good ad, or the best ad, we want to show it often so that we maximize expected revenue

- Problem: what is the optimal strategy for showing the K ads?
  - Strategy = sequence of (ad, click/no-click) pairs

The Multi-Armed Bandit Problem

- Model the explore/exploit problem as a "multi-armed bandit", i.e., as a slot machine for gambling with K arms
- Each "arm" corresponds to an ad, with "payoff" probability p_k , $k = 1,\ldots K$
  - Assume for simplicity that if we pull an arm and "win" we get rewarded 1 unit
- Objective: construct N successive pulls of the slot machine to maximize the expected total reward
- This is a well-studied problem in sequential optimization
  - e.g., Asymptotically efficient adaptive allocation rules, Lai and Robbins, *Advances in Applied Mathematics*, 6:4-22, 1985
  - Even earlier work dating back to the 1950's
  - Other instances of this problem occur in applications where you have to make choices "along the way" from a finite set of options based only on partial information

Theoretical Framework

- K bandits, with payoff probabilities p_k , $k = 1,....K$, and unit rewards = 1
  - Assume for simplicity that p_k probabilities and rewards don't change over time
  - Also assume that bandits are memoryless (as in coin-tossing)

- Let X_k be the reward on any trial for bandit k. Assume for simplicity that

  X_k =1 with probability p_k, and = 0 with probability 1 - p_k

  Expected reward from bandit k is $E[X_k] = 1\, p_k + 0\,(1- p_k) = p_k$

- Optimal strategy to maximize the expected reward?
  - Always select the k value that maximizes $E[X_k]$, i.e., the largest probability p_k
  - This optimal strategy exists only in theory, if we know the p_k 's (which we don't)

- Various theoretical analyses look at what happens on average by using certain types of strategies.

  Expected Regret(S) = E [reward | optimal strategy] – E [reward | strategy S]

Naïve Strategies

- Deterministic Greedy Strategy:
  - at iteration N, pick the bandit that has performed best up to this time
  - Weakness?
    - Will under-explore bandits and may easily select a sub-optimal bandit forever

- Play-the-Winner Strategy
  - At iteration N
    - play the bandit from iteration N-1 if it was successful, otherwise
    - select another arm uniformly at random or cycle through them deterministically
  - This is the optimal thing to do if the bandit was successful at time N-1
  - But not necessarily optimal to switch away from this bandit if it failed
  - Thus, this strategy tends to switch too much and over-explores

    (see Berry and Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Chapman & Hall, 1985)

Note that both strategies above perform even more poorly if the learning is happening in batch mode rather than at each iteration.

Simple Example of Multi-Armed Bandit Strategy

- Epsilon-Greedy Strategy
  - At iteration t in the algorithm
  - Select the best bandit (up to this point) with probability, $1 - \varepsilon$, e.g., $\varepsilon = 0.1$
  - Select one of the other K-1 bandits with probability ε
    - uniformly at random
    - or in proportion to their estimated p_k at this point

- Key aspects of the strategy
  - How to select ε
    - If its too small, we won't explore enough
    - If its too large, we won't exploit enough
  - How do we define "best"?
    - E.g., raw frequency $p_k = r_k / N_k$, or a smoothed estimate?

- Weakness?
  - ε is fixed: so it continues to explore with probability ε, long after the best bandit has been identified – and hence is suboptimal

Other Examples of Strategies

- Epsilon-greedy where we decrease ε as the experiment progress
  - Makes intuitive sense: explore a lot at first, then start to exploit more
  - Adds an additional "tuning" parameter of how to decrease ε

- Epsilon-first Strategy
  - Pure exploration followed by pure exploitation
  - First explore for εN trials, selecting bandits uniformly at random
  - Then exploit for $(1-\varepsilon)N$ trials, selecting the best bandit from the explore phase

- Theoretical analyses provide results like bounds on the rates at which arms should be played, as a function of the true (unknown) p_k values
  - These results provide very useful insights and general guidance
  - But don't provide specific strategies

Randomized Probability Matching Strategy

- Idea: number of pulls from bandit k should be proportional to the probability that bandit k is optimal
  - Also known as Thompson sampling or “Bayesian bandits”

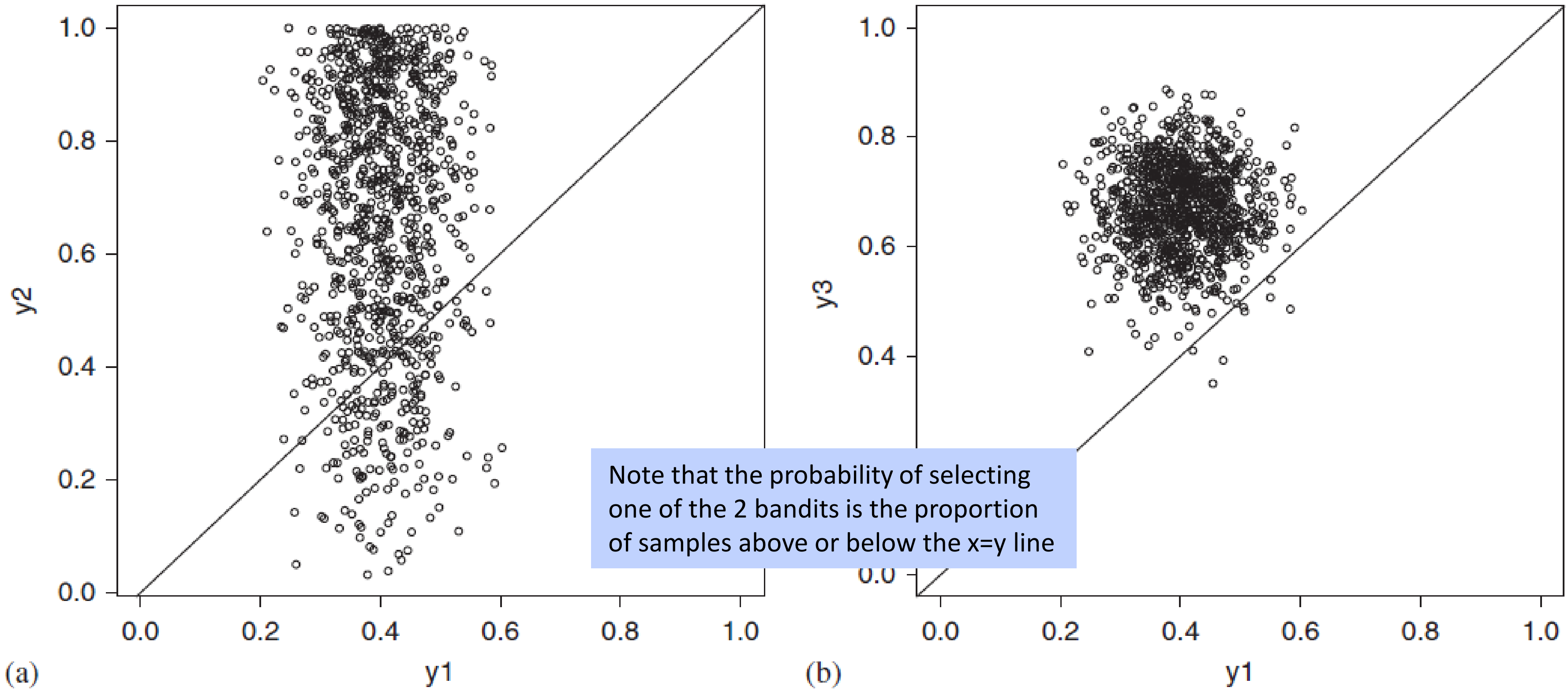
- Let $P(p_k \mid r_k, N_k)$ be a Bayesian density on the value p_k
  - where r_k, N_k = number of trials and successes with the kth bandit so far
  - $P(p_k \mid r_k, N_k)$ is our posterior belief about p_k, given the data r_k, N_k
  - e.g., using a Beta prior and a Beta posterior density

- At each iteration we do the following:
  - Sample M values of p_k for each bandit k from its density $P(p_k \mid r_k, N_k)$
  - For each bandit compute w_k = proportion of M samples that bandit k has the largest p_k value
  - Select a bandit k by sampling from the distribution $w = [w_1, \ldots, w_K]$
  - Update the r_k, N_k values and update the density $P(p_k \mid r_k, N_k)$

Simulation example showing 1000 draws from posterior distributions on bandit probabilities

Y-axis: 2 successes, 1 Failure to date
X-axis: 20 successes, 30 Failures to date

Y-axis: 20 successes, 10 Failures to date
X-axis: 20 successes, 30 Failures to date



Note that the probability of selecting one of the 2 bandits is the proportion of samples above or below the x=y line

Figure 1. One thousand draws from the joint distribution of two independent beta distributions. In both cases, the horizontal axis represents a beta (20,30) distribution. The vertical axis is (a) beta(2,1) and (b) beta(20,10).

Figure from S. L. Scott, A modern Bayesian look at the multi-armed bandit, *Applied Stochastic Models in Business and Industry*, 26:639-658, 2010

Randomized Probability Matching Strategy

- Strengths
  - Works well on a wide-range of problems
  - Relatively simple to implement
  - Relatively free of tuning parameters
  - Flexible enough to accommodate more complicated versions of the problem
  - Balances exploration and exploitation in an intuitive way

- Weaknesses
  - Requires more computation to select an arm at each iteration
  - Theoretical results/guarantees, relative to other methods, not generally known (yet)

For additional discussion and experiments see S. L. Scott, A modern Bayesian look at the multi-armed bandit, *Applied Stochastic Models in Business and Industry*, 26:639-658, 2010

Click Fraud

- Click fraud = generation of artificial (non-human) clicks for ads

- Why?
  - Artificially increases the costs for the advertiser (for CPC)
  - Artificially increases the revenue of the site hosting the ad (for CPC)

- Click Quality Teams
  - All major search engines have full-time teams monitoring/managing click fraud
  - Use a combination of human analysis and machine learning algorithms

- Controversial topic
  - Advertisers say search engines are not doing enough, claim fraud clicks are > 20%
  - Search engines reluctant to publish too much data on frauds, claim fraud click percentage is much lpower

Advertising on the Web

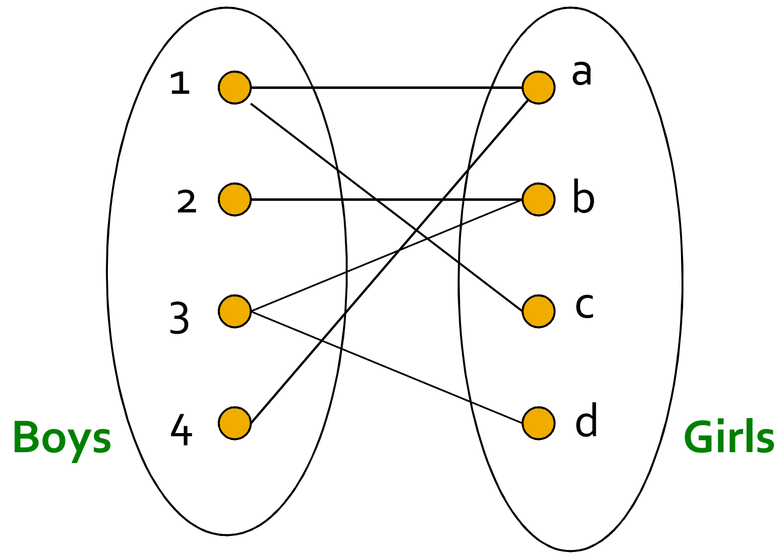
Mining of Massive Datasets
Jure Leskovec, Anand Rajaraman, Jeff Ullman
Stanford University
http://www.mmds.org

Online Algorithms

- **Classic model of algorithms**
  - You get to see the entire input, then compute some function of it
  - In this context, "offline algorithm"
- **Online Algorithms**
  - You get to see the input one piece at a time, and need to make irrevocable decisions along the way
  - **Similar to the data stream model**

Online Bipartite Matching

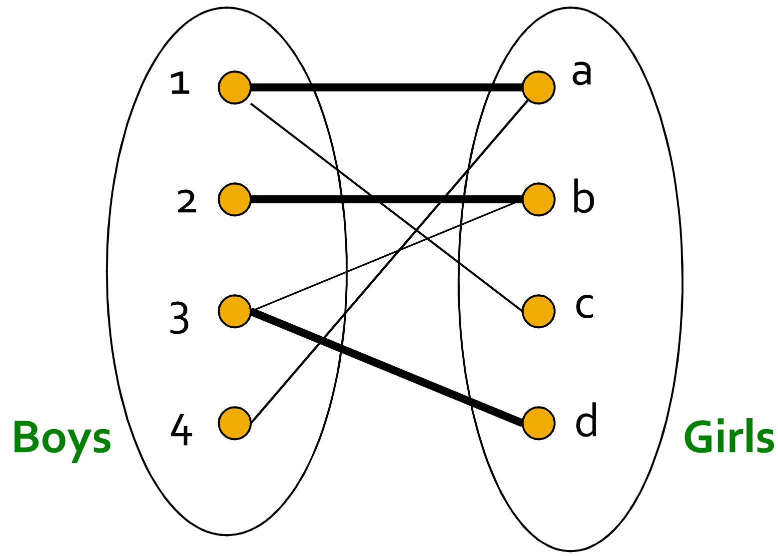
Example: Bipartite Matching



Nodes: Boys and Girls; Edges: Preferences

Goal: Match boys to girls so that maximum number of preferences is satisfied

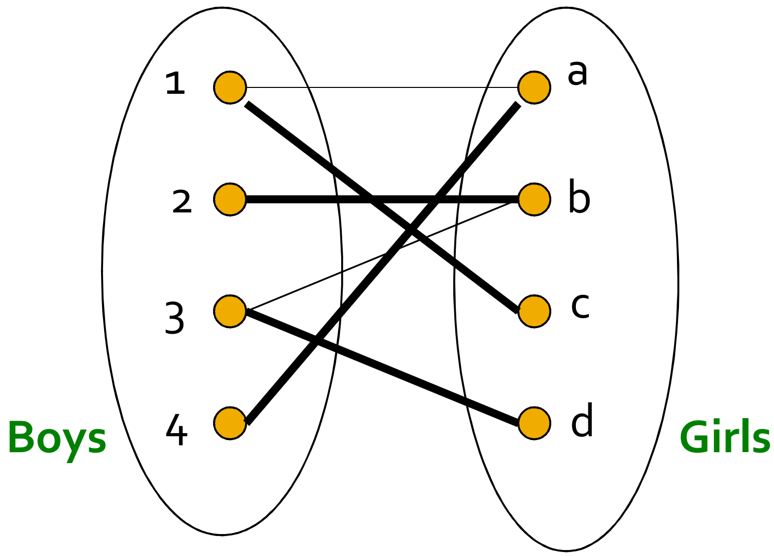
Example: Bipartite Matching



M = {(1,a),(2,b),(3,d)} is a **matching**

Cardinality of matching = |M| = 3

Example: Bipartite Matching



M = {(1,c),(2,b),(3,d),(4,a)} is a
perfect matching

Perfect matching … all vertices of the graph are matched
Maximum matching … a matching that contains the largest possible number of matches

Matching Algorithm

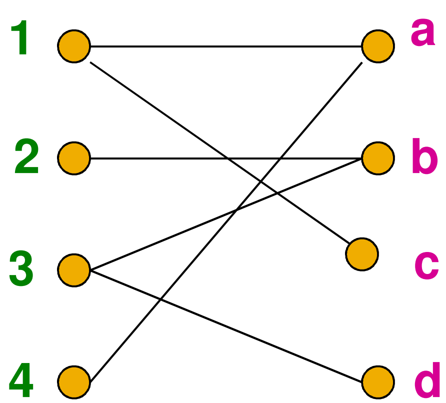
- **Problem: Find a maximum matching for a given bipartite graph**
  - A perfect one if it exists
- There is a polynomial-time offline algorithm based on augmenting paths (Hopcroft & Karp 1973, see http://en.wikipedia.org/wiki/Hopcroft-Karp_algorithm)
- **But what if we do not know the entire graph upfront?**

Online Graph Matching Problem

- Initially, we are given the set **boys**
- In each **round**, **one girl's choices are revealed**
  - That is, girl's **edges** are revealed
- **At that time, we have to decide to either:**
  - Pair the **girl** with a **boy**
  - Do not pair the **girl** with any **boy**

- **Example of application:**
  Assigning tasks to servers

Online Graph Matching: Example



(1,a)
(2,b)
(3,d)

Greedy Algorithm

- **Greedy algorithm for the online graph matching problem:**
  - Pair the new girl with **any** eligible boy
    - If there is none, do not pair girl
- **How good is the algorithm?**

Competitive Ratio

- For input ***I***, suppose greedy produces matching $\boldsymbol{M_{greedy}}$ while an optimal matching is $\boldsymbol{M_{opt}}$

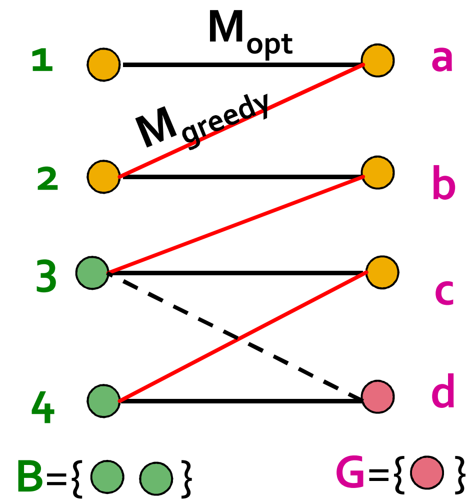
Competitive ratio =

$$\min_{\text{all possible inputs } I} \left(|M_{greedy}| / |M_{opt}|\right)$$

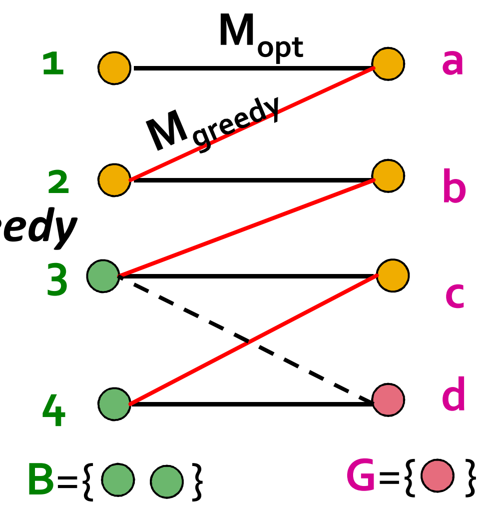
(what is greedy's <u>worst</u> performance <u>over all possible</u> inputs *I*)

Analyzing the Greedy Algorithm

- Consider a case: $M_{greedy} \neq M_{opt}$
- Consider the set G of girls matched in M_{opt} but not in M_{greedy}
- Then every boy B adjacent to girls in G is already matched in M_{greedy}:
  - If there would exist such non-matched (by M_{greedy}) boy adjacent to a non-matched girl then greedy would have matched them
- Since boys B are already matched in M_{greedy} then
  (1) $|M_{greedy}| \geq |B|$

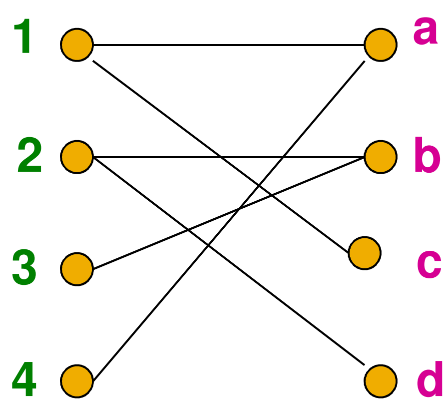


Analyzing the Greedy Algorithm



- **Summary so far:**
  - Girls G matched in M_{opt} but not in M_{greedy}
  - **(1)** $|M_{greedy}| \geq |B|$
- There are at least $|G|$ such boys ($|G| \leq |B|$) otherwise the optimal algorithm couldn't have matched all girls in G
  - **So:** $|G| \leq |B| \leq |M_{greedy}|$
- By definition of G also: $|M_{opt}| \leq |M_{greedy}| + |G|$
  - Worst case is when $|G| = |B| = |M_{greedy}|$
- $|M_{opt}| \leq 2|M_{greedy}|$ **then** $|M_{greedy}|/|M_{opt}| \geq 1/2$

Worst-case Scenario



(1,a)
(2,b)

Web Advertising

History of Web Advertising

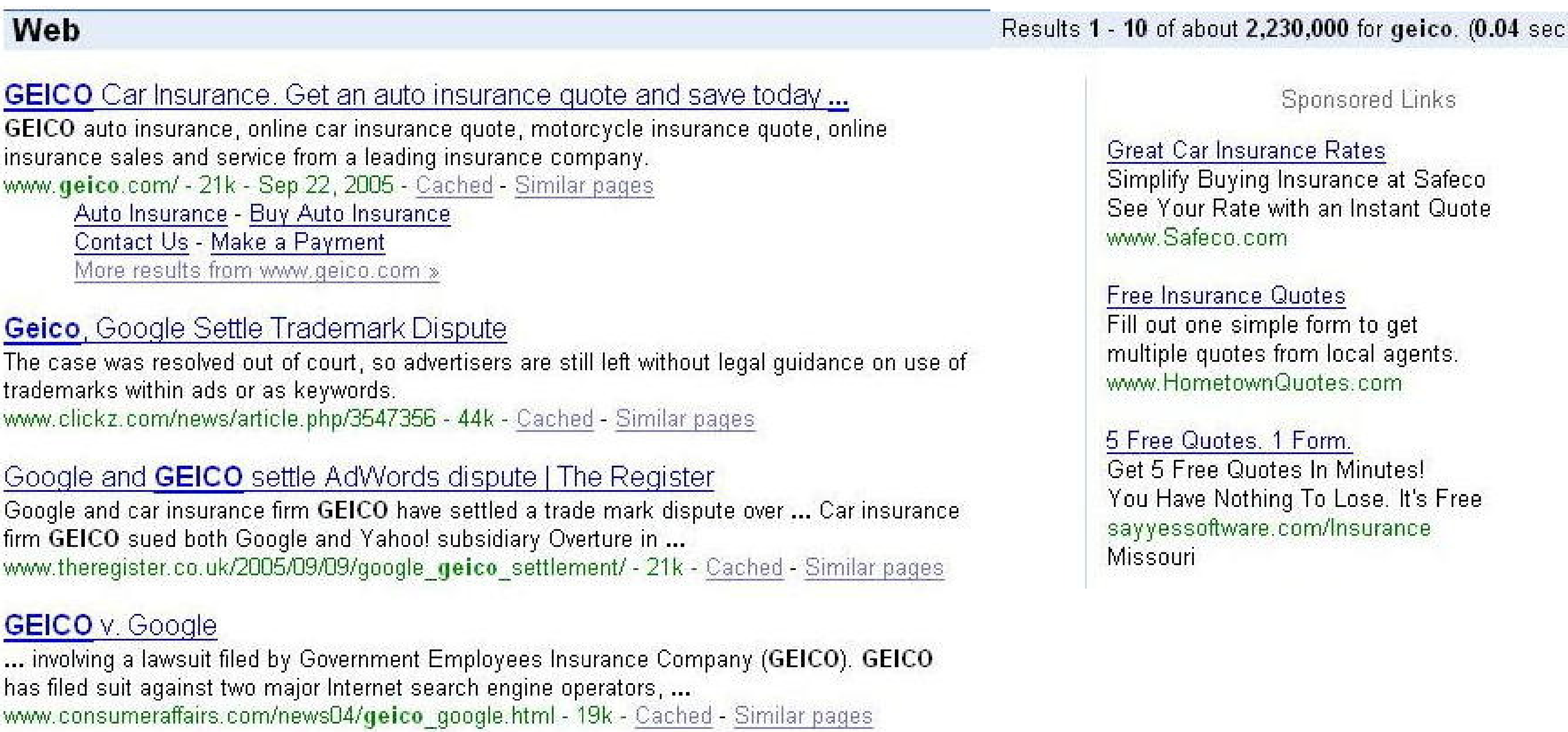
- **Banner ads (1995-2001)**
  - Initial form of web advertising
  - Popular websites charged *X*$ for every 1,000 "impressions" of the ad
    - Called "**CPM**" rate (Cost per thousand impressions)
    - Modeled similar to TV, magazine ads
  - From **untargeted** to **demographically targeted**
  - **Low click-through rates**
    - Low ROI for advertisers

CPM…cost per *mille*
Mille…thousand in Latin

Performance-based Advertising

- **Introduced by Overture around 2000**
  - Advertisers **bid** on **search keywords**
  - When someone searches for that keyword, the **highest bidder's ad is shown**
  - Advertiser is charged only if the ad is clicked on
- Similar model adopted by Google with some changes around 2002
  - Called **Adwords**

Ads vs. Search Results



Web

Results **1** - **10** of about **2,230,000** for **geico**. (**0.04** secc

GEICO Car Insurance. Get an auto insurance quote and save today ...
GEICO auto insurance, online car insurance quote, motorcycle insurance quote, online insurance sales and service from a leading insurance company.
www.**geico**.com/ - 21k - Sep 22, 2005 - Cached - Similar pages
Auto Insurance - Buy Auto Insurance
Contact Us - Make a Payment
More results from www.geico.com »

Geico, Google Settle Trademark Dispute
The case was resolved out of court, so advertisers are still left without legal guidance on use of trademarks within ads or as keywords.
www.clickz.com/news/article.php/3547356 - 44k - Cached - Similar pages

Google and GEICO settle AdWords dispute | The Register
Google and car insurance firm **GEICO** have settled a trade mark dispute over **...** Car insurance firm **GEICO** sued both Google and Yahoo! subsidiary Overture in **...**
www.theregister.co.uk/2005/09/09/google_**geico**_settlement/ - 21k - Cached - Similar pages

GEICO v. Google
... involving a lawsuit filed by Government Employees Insurance Company (**GEICO**). **GEICO** has filed suit against two major Internet search engine operators, **...**
www.consumeraffairs.com/news04/**geico**_google.html - 19k - Cached - Similar pages

Sponsored Links

Great Car Insurance Rates
Simplify Buying Insurance at Safeco
See Your Rate with an Instant Quote
www.Safeco.com

Free Insurance Quotes
Fill out one simple form to get
multiple quotes from local agents.
www.HometownQuotes.com

5 Free Quotes. 1 Form.
Get 5 Free Quotes In Minutes!
You Have Nothing To Lose. It's Free
sayyessoftware.com/Insurance
Missouri

Web 2.0

- **Performance-based advertising works!**
  - Multi-billion-dollar industry

- **Interesting problem:**
  What ads to show for a given query?
  - (Today's lecture)

- **If I am an advertiser, which search terms should I bid on and how much should I bid?**
  - (Not focus of today's lecture)

Adwords Problem

- **Given:**
  - **1.** A set of bids by advertisers for search queries
  - **2.** A click-through rate for each advertiser-query pair
  - **3.** A budget for each advertiser (say for 1 month)
  - **4.** A limit on the number of ads to be displayed with each search query
- **Respond to each search query with a set of advertisers such that:**
  - **1.** The size of the set is no larger than the limit on the number of ads per query
  - **2.** Each advertiser has bid on the search query
  - **3.** Each advertiser has enough budget left to pay for the ad if it is clicked upon

Adwords Problem

- A stream of queries arrives at the search engine: $q_1, q_2, \ldots$
- Several advertisers bid on each query
- When query q_i arrives, search engine must pick a subset of advertisers whose ads are shown

- **Goal: Maximize search engine's revenues**
  - **Simple solution:** Instead of raw bids, use the "**expected revenue per click**" (i.e., **Bid*CTR**)
- **Clearly we need an online algorithm!**

The Adwords Innovation

Advertiser	Bid	CTR	Bid * CTR
A	$1.00	1%	1 cent
B	$0.75	2%	1.5 cents
C	$0.50	2.5%	1.125 cents

Click through rate

Expected revenue

The Adwords Innovation

Advertiser	Bid	CTR	Bid * CTR
B	$0.75	2%	1.5 cents
C	$0.50	2.5%	1.125 cents
A	$1.00	1%	1 cent

Complications: Budget

- **Two complications:**
  - **Budget**
  - **CTR of an ad is unknown**

- **Each advertiser has a limited budget**
  - **Search engine guarantees that the advertiser will not be charged more than their daily budget**

Complications: CTR

- **CTR: Each ad has a different likelihood of being clicked**
  - **Advertiser 1** bids $2, click probability = 0.1
  - **Advertiser 2** bids $1, click probability = 0.5
  - **Clickthrough rate (CTR)** is measured **historically**
    - **Very hard problem: Exploration vs. exploitation**
      Exploit: Should we keep showing an ad for which we have good estimates of click-through rate
      or
      Explore: Shall we show a brand new ad to get a better sense of its click-through rate

Greedy Algorithm

- **Our setting: Simplified environment**
  - There is **1** ad shown for each query
  - All advertisers have the same budget ***B***
  - All ads are equally likely to be clicked
  - Value of each ad is the same (=**1**)
- **Simplest algorithm is greedy:**
  - For a query pick any advertiser who has bid **1** for that query
  - **Competitive ratio of greedy is 1/2**

Bad Scenario for Greedy

- **Two advertisers A and B**
  - ***A*** bids on query ***x***, ***B*** bids on ***x*** and ***y***
  - Both have budgets of **$4**
- **Query stream:** ***x x x x y y y y***
  - Worst case greedy choice: ***B B B B*** **_ _ _ _**
  - Optimal: **A A A A B B B B**
  - **Competitive ratio = ½**
- **This is the worst case!**
  - **Note:** Greedy algorithm is deterministic – it always resolves draws in the same way

BALANCE Algorithm [MSVV]

- **BALANCE** Algorithm by Mehta, Saberi, Vazirani, and Vazirani
  - **For each query, pick the advertiser with the largest unspent budget**
    - Break ties arbitrarily (**but in a deterministic way**)

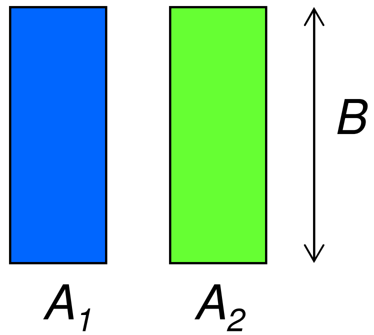
Example: BALANCE

- **Two advertisers A and B**
  - **A** bids on query ***x***, **B** bids on ***x*** and ***y***
  - Both have budgets of **$4**
- **Query stream:** ***x x x x y y y y***
- **BALANCE choice:** **A B A B B B _ _**
  - Optimal: **A A A A B B B B**
- **In general:** For **BALANCE** on **2** advertisers
  Competitive ratio = ¾

Analyzing BALANCE

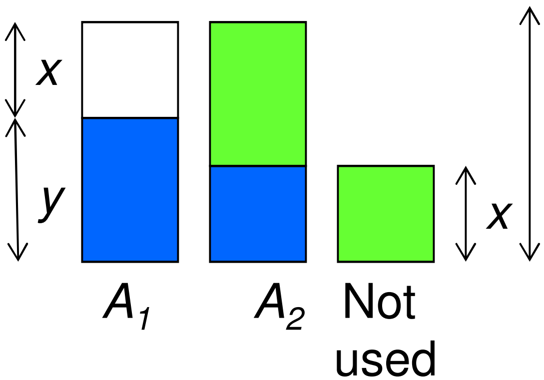
- **Consider simple case (w.l.o.g.):**
  - **2** advertisers, $\mathbf{A_1}$ and $\mathbf{A_2}$, each with budget **B** ($\geq$1)
  - Optimal solution exhausts both advertisers' budgets
- **BALANCE must exhaust at least one advertiser's budget:**
  - **If not, we can allocate more queries**
    - Whenever BALANCE makes a mistake (both advertisers bid on the query), advertiser's unspent budget only decreases
    - Since optimal exhausts both budgets, one will for sure get exhausted
  - Assume BALANCE exhausts A_2's budget, but allocates x queries fewer than the optimal
  - **Revenue:** $BAL = 2B - x$

Analyzing Balance



Queries allocated to A_1 in the optimal solution

Queries allocated to A_2 in the optimal solution



Optimal revenue = **2B**
Assume Balance gives revenue **= 2B-x = B+y**

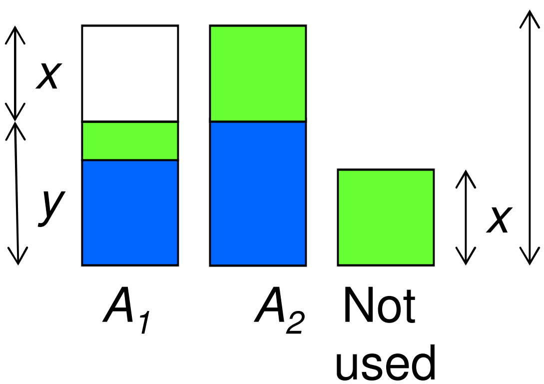
Unassigned queries should be assigned to A_2
(if we could assign to A_1 we would since we still have the budget)

Goal: Show we have **$y \geq x$**

Case 1) ≤ ½ of A_1's queries got assigned to A_2
then $y \geq B/2$

Case 2) > ½ of A_1's queries got assigned to A_2
then $x \leq B/2$ **and** $x + y = B$

Balance revenue is minimum for $x = y = B/2$
Minimum Balance revenue = $3B/2$
Competitive Ratio = 3/4



BALANCE exhausts A_2's budget

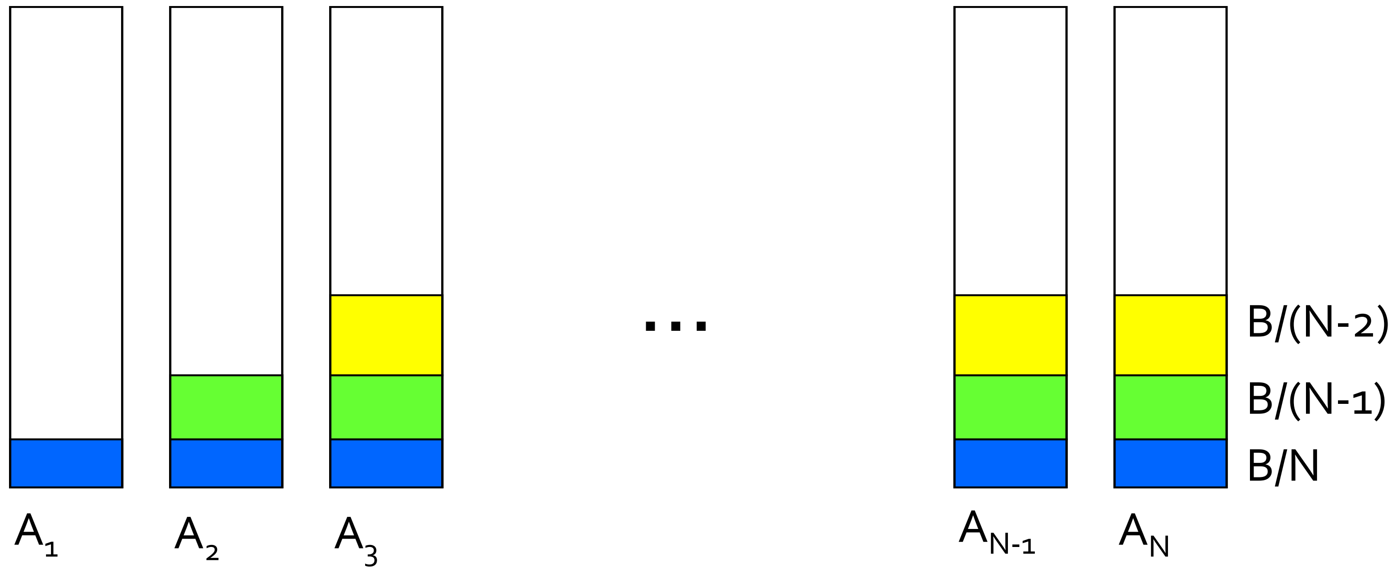
BALANCE: General Result

- **In the general case, worst competitive ratio of BALANCE is $1 – 1/e$ = approx. 0.63**
  - Interestingly, no online algorithm has a better competitive ratio!
- **Let's see the worst case example that gives this ratio**

Worst case for BALANCE

- ***N* advertisers: $A_1, A_2, \ldots A_N$**
  - Each with budget $B > N$
- **Queries:**
  - $N \cdot B$ queries appear in ***N*** rounds of ***B*** queries each
- **Bidding:**
  - Round **1** queries: bidders $\mathbf{A_1, A_2, \ldots, A_N}$
  - Round **2** queries: bidders $\mathbf{A_2, A_3, \ldots, A_N}$
  - Round ***i*** queries: bidders $\mathbf{A_i, \ldots, A_N}$
- **Optimum allocation:**
  Allocate round ***i*** queries to A_i
  - Optimum revenue $N \cdot B$

BALANCE Allocation



BALANCE assigns each of the queries in round 1 to **N** advertisers.
After ***k*** rounds, sum of allocations to each of advertisers $\boldsymbol{A_k,\dots,A_N}$ is

$$\boldsymbol{S_k = S_{k+1} = \cdots = S_N = \sum_{i=1}^{k-1} \frac{B}{N-(i-1)}}$$

If we find the smallest *k* such that $S_k \geq B$, then after *k* rounds we cannot allocate any queries to any advertiser

BALANCE: Analysis

$$B/1 \quad B/2 \quad B/3 \quad \ldots \quad B/(N-(k-1)) \ \ldots \ B/(N-1) \quad B/N$$

S_1

S_2

S_k **= B**

$$1/1 \quad 1/2 \quad 1/3 \quad \ldots \quad 1/(N-(k-1)) \ \ldots \ 1/(N-1) \quad 1/N$$

S_1

S_2

S_k **= 1**

BALANCE: Analysis

- **Fact:** $H_n = \sum_{i=1}^{n} 1/i \approx \ln(n)$ for large n
  - Result due to Euler

1/1 1/2 1/3 … 1/(N-(k-1)) … 1/(N-1) 1/N

ln(N)

ln(N)-1 $S_k = 1$

- $S_k = 1$ implies: $H_{N-k} = ln(N) - 1 = ln(\frac{N}{e})$
- We also know: $H_{N-k} = ln(N - k)$
- So: $N - k = \frac{N}{e}$
- **Then:** $k = N(1 - \frac{1}{e})$

N terms sum to ln(*N*).
Last *k* terms sum to 1.
First *N*-*k* terms sum
to ln(*N*-*k*) but also to ln(*N*)-1

BALANCE: Analysis

- So after the first **k=N(1-1/e)** rounds, we cannot allocate a query to any advertiser
- **Revenue = B·N (1-1/e)**
- **Competitive ratio = 1-1/e**

General Version of the Problem

- **Arbitrary bids and arbitrary budgets!**
- Consider we have 1 query q, advertiser i
  - Bid = x_i
  - Budget = b_i
- **In a general setting BALANCE can be terrible**
  - Consider two advertisers A_1 and A_2
  - A_1: $x_1 = 1$, $b_1 = 110$
  - A_2: $x_2 = 10$, $b_2 = 100$
  - Consider we see **10** instances of **q**
  - BALANCE always selects A_1 and earns **10**
  - Optimal earns **100**

Generalized BALANCE

- **Arbitrary bids:** consider query q, bidder i
  - Bid = x_i
  - Budget = b_i
  - Amount spent so far = m_i
  - Fraction of budget left over $f_i = 1-m_i/b_i$
  - Define $\psi_i(q) = x_i(1-e^{-f_i})$
- Allocate query q to bidder i with largest value of $\psi_i(q)$
- **Same competitive ratio (1-1/e)**