

(A Few) Key Lessons Learned Building Recommender Systems For Large-scale Social Networks



Christian Posse



Principal Data Scientist and Product Manager at
LinkedIn

San Francisco Bay Area | Information Technology and Services

World's Largest Professional Network

175M+

62% non U.S.



2/sec



25th

Most visit website worldwide
(Comscore 6-12)



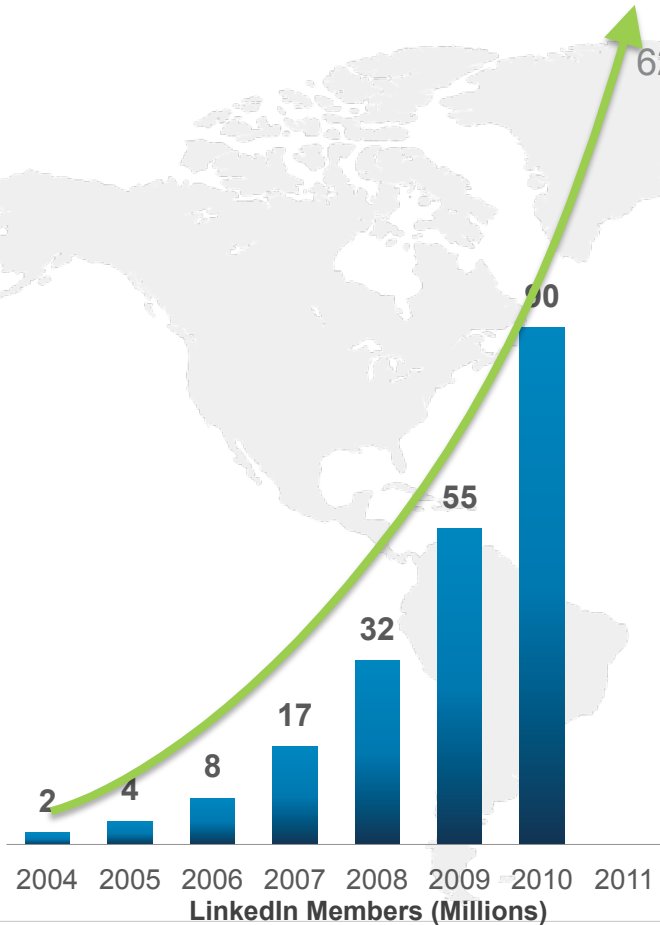
>2M

Company pages



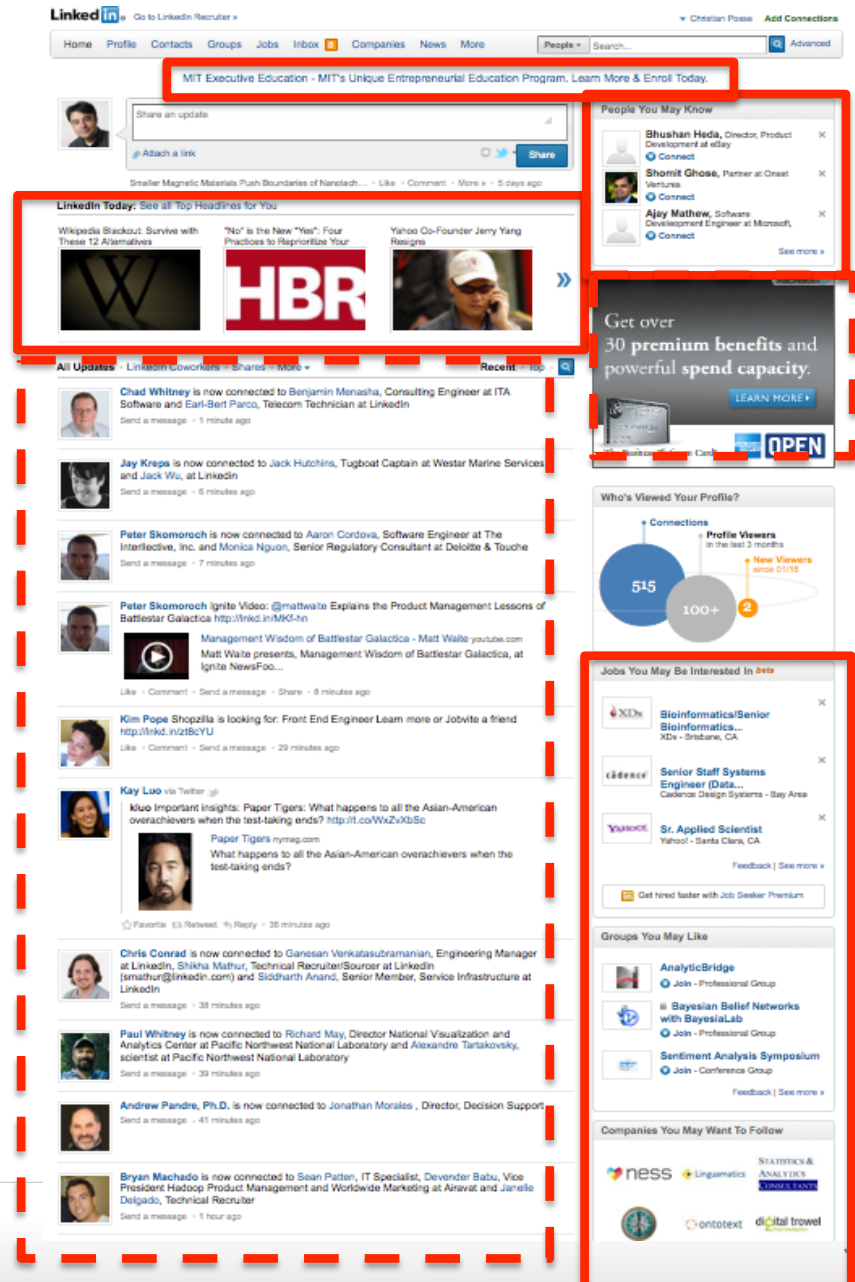
85%

Fortune 500 Companies use
LinkedIn to hire



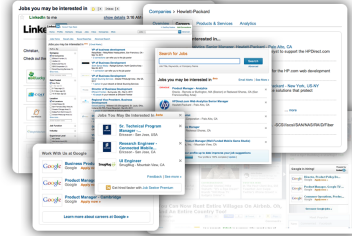
LinkedIn Homepage

Powered by
Recommendations!

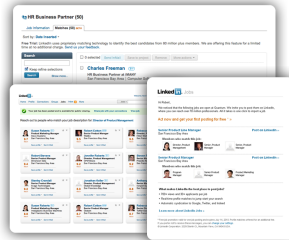


The Recommendations Opportunity

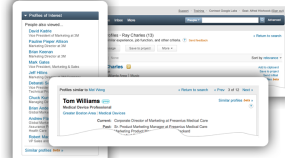
Jobs You May Be Interested In



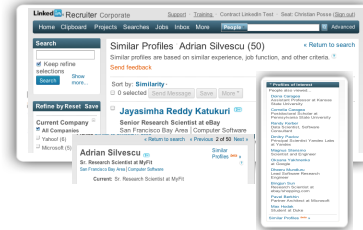
Talent Match



CAP

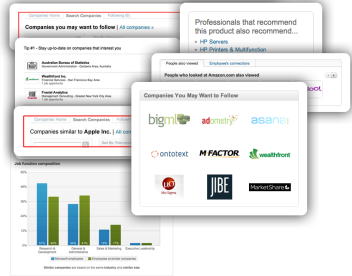


Similar Profiles

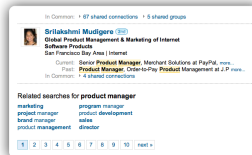


Companies

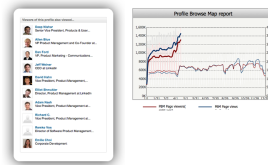
Recommendations, similar companies search, peer companies, and company browse maps, company products and services browse maps



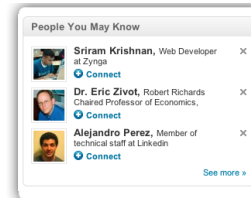
Related search



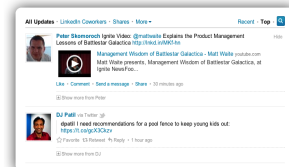
Profile browse maps



Connections



Network updates



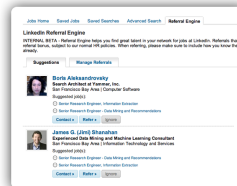
Jobs browse maps



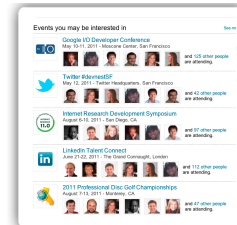
Ad matching engine

$pCTR = \frac{(member, creative, advertiser, context, inventory, OCTR)}{...}$

Referral Engine

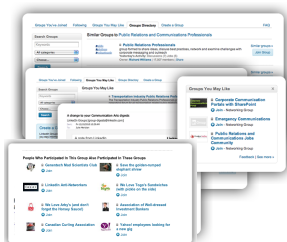


Events You May Be Interested In

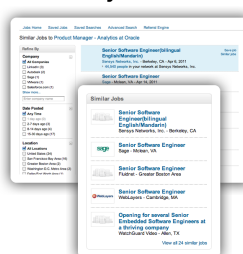


Groups

Recommendations, similar groups search



Similar jobs



News

LinkedIn Today: See all Top Headlines for You



What are they worth? Think 50%

- > 50% of connections are from recommendations (PYMK)
- > 50% of job applications are from recommendations (JYMBII)
- > 50% of group joins are from recommendations (GYML)

People You May Know *updated*



Sophie Perisic, French Teacher at Crystal Springs Uplands School
[Connect](#)



Lissa Juan, Senior Manager, Product Management - Mobile &
[Connect](#)



Jeff King, Sr. Director X.commerce Platform Partnerships at eBay
[Connect](#)

[See more »](#)

Jobs You May Be Interested In *beta*



Director of Engineering
Pandora - Oakland, CA



Director of Engineering for Core...
CyberCoders - Santa Clara, CA



Director of Engineering
HotelTonight - San Francisco Bay...

[Feedback](#) | [See more »](#)

Groups You May Like



LinkedIn Groups Product Forum
[Join - Other...](#)



RDataMining
[Join - Professional Group](#)



Data Mining, Statistics, and Data Visualization
[Join - Professional Group](#)

[Feedback](#) | [See more »](#)

What is a Recommender System?

A Recommender selects a product that if acquired by the “buyer” maximizes value of both “buyer” and “seller” at a given point in time

Lesson 1 : Recommendations must make strategic sense...

- Conflicts of interest between 'buyers'
 - E.g. job posters vs. job seekers



Director, Data Scientist
PayPal - San Jose, CA (San Francisco Bay Area)



Job Description

PayPal is the faster, safer way to pay and get paid. The service allows members to send money without sharing financial information, with the flexibility to pay using their account balances, bank accounts, credit cards or promotional financing. With nearly 100 million active accounts in 190 markets and 19 currencies around the world, PayPal enables global commerce. Putting customers' needs first, PayPal is redefining the payments category through product, marketing, and service delivery innovation. We're looking for engineers, applied scientists, and researchers with experience in implementing data mining and machine learning systems to help build innovative products all of PayPal and beyond. This is a new team under the guidance of the chief scientist at PayPal that will leverage all data throughout eBay Inc. -- from creating new products, technologies and platforms, all the way to supplement existing products for various use cases. We hope to build a world-class data science team that will have direct impact on

Apply on Company Website

Save job

Share job   

Stop following

Posted By

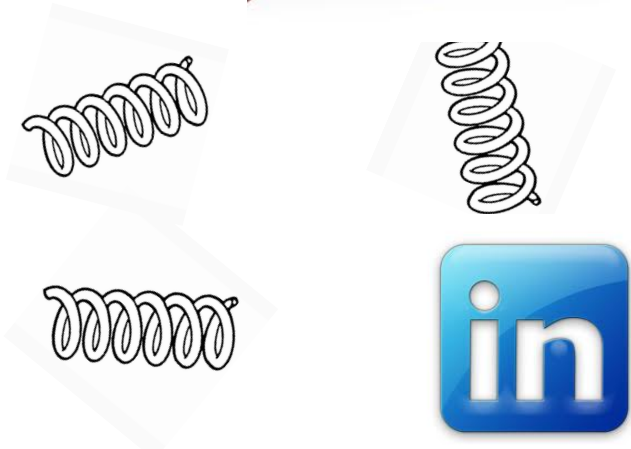


Sachi Yokota (2nd)
Recruiter, Marketing & Strategy at PayPal
[Send InMail](#)

1 of your connections can refer you to Sachi:



John Reilly (1st)



- What is the (long-term) value of an action?
 - How do we compare an invitation from a job application from a group join or the reading of a news?

Ingredients of a Recommender System

A Recommender processes information and transforms it into actionable knowledge



Lesson 2: User Experience Matters Most

1. Understand **user intent**
→ Define, model, leverage
2. Be in the **user flow**
3. Optimize **location** on the page
4. Set **right** expectations (“You May...”)
5. **Explain** recommendations
6. **Interact** with the user
7. Leverage **Social Referral**

User Intent, User Flow, Location, Message

Job recommendation use cases

User experience	Optimization	Impact on job application rate
User intent / location	Homepage personalization	2.5X
User flow / user intent	Before vs. after having applied to a job	7X
User flow	LinkedIn homepage vs. Jobs homepage	10X
Location	Center rail vs. right rail	5X
Message	Followers vs. leaders	2X

Applied Researcher / Data Scientist: Data Analysis, Data Mining and Machine Learning
 LinkedIn - Mountain View, CA (San Francisco Bay Area)

Job Description
 The engineering culture at LinkedIn is based on building and integrating cutting-edge technologies while encouraging creativity, innovation and expansion. Our engineers constantly raise the bar for excellence, motivating each other to tackle challenges and take intelligent risks. The industry is moving fast and our engineers are right there with it!
<http://engineering.linkedin.com/>

About Data Analysis, Data Mining and Machine Learning
 We're looking for an Applied Researcher / Scientist to work on our massive semi-structured text and graph datasets. This role is ideal for a senior researcher who has both domain experience in data mining, information retrieval, or machine learning and a strong systems orientation. It offers the unique opportunity to conduct applied research and have a huge end-to-end impact on key LinkedIn product initiatives, including search relevance, ad targeting, information extraction, and recommendations.

Desired Skills & Experience

- You are a computer science champ with strong coding ability and familiarity with Java, C++ or any other OOP language.
- You have worked with data mining toolkits like Weka, Mahout, R, NLTK, etc.
- You have experience with Hadoop or other MapReduce paradigms.
- You're familiar with information retrieval libraries like Lucene / SOLR.
- You're interested in big data, crunching billions of examples for statistical modeling, data mining, recommendation or search relevance solutions.
- You have published in academic and industry circles.
- BS, MS, or PhD in computer science or other quantitative discipline.
- You thrive in a fast paced, fast-driven, collaborative and iterative programming environment.
- You enjoy using data to drive products.

Company Description
 Founded in 2003, LinkedIn connects the world's professionals to make them more productive and successful. With more than 130 million members worldwide, including executives from every Fortune 500 company, LinkedIn is the world's largest professional network online. The company has a diversified business model with revenues coming from member subscriptions, marketing solutions and hiring solutions. Headquartered in Silicon Valley, LinkedIn has offices in 22 countries across the globe.

Additional Information
 Posted: November 28, 2011
 Type: Full-time
 Experience: Mid-Senior level
 Functions: Engineering, Research, Science
 Industries: Internet, Information Technology and Services, Computer Software
 Job ID: 1936094

People Who Viewed This Job Also Viewed

- Data Scientist at LivingSocial
- Research Scientist at Yahoo!
- Data Mining Consultant at IBM
- Software Engineer, Applications at LinkedIn
- Senior Algorithm/Software Engineer (C/C++ / Classification Algorithms / Image Processing / Machine Learning) at KLU-Tensor
- Artificial Intelligence and Reasoning Engineer - Corporate R&D at Qualcomm
- Machine Learning Software Engineer, YouTube - Mountain View at Google
- Data Scientist (PL) at LinkedIn

Similar Jobs

- Data Scientist at LivingSocial - Washington D.C. Metro Area
- Graduate Summer Internship in Machine Learning at Intel - Redwood, MA
- Software Engineer, Machine Learning at Oracle - Mountain View, CA
- Data Scientist at Verta - Cambridge, MA
- Principal Software Engineer at LinkedIn - Mountain View, CA

Item based collaborative filtering:
 → Follower audience

Content based:
 → Leader audience

User Intent Modeling

1. Intent labeling

- Job seeker, recruiter, outbound professional, content consumer, content creator, networker, profile curator

2. Build normalized propensity model for each intent

Job Seeker Intent

- Features

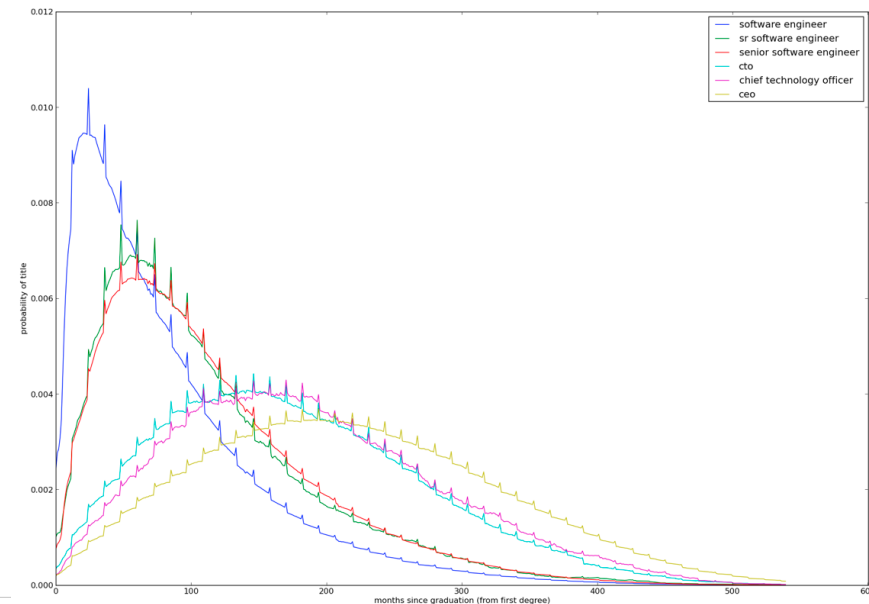
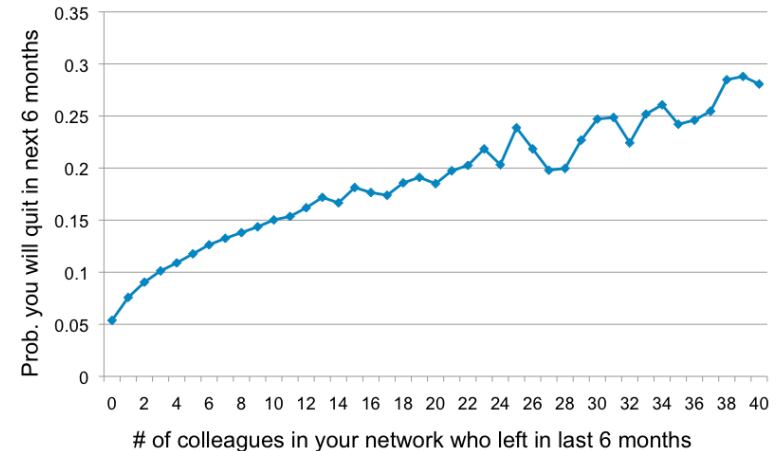
- Behavior: job searches, views & applications, job related email replies, has a job seeker sub, ...
- Social graphs: # of colleagues in network who recently left, ...
- Content: title, industry, anniversary date, ...

- Propensity Score

- Parametric accelerated failure time survival model
 - $\log T_i = \sum_k \beta_k x_{ik} + \sigma \varepsilon_i$
- Score: P(switch job next month)

- Evaluation

- Gold standard
- Directional validation

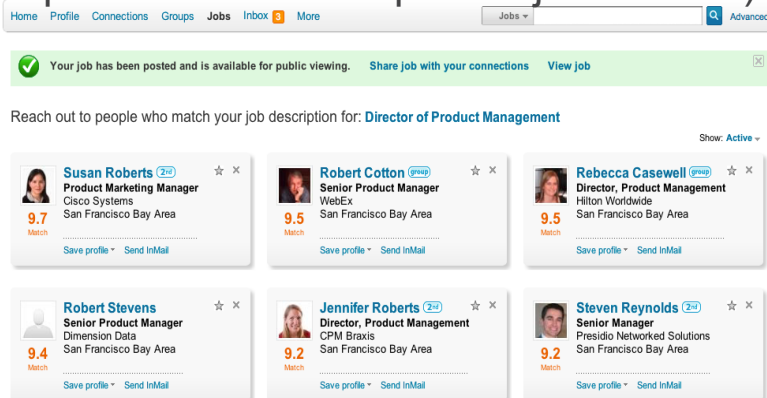


Lesson 3: Recommendations often cater to multiple competing objectives



Suggest relevant groups ... that one is more likely to participate in

TalentMatch (Top 24 matches of a posted job for sale)



Suggest skilled candidates ... who will likely respond to hiring managers inquiries

Semantic + engagement objectives

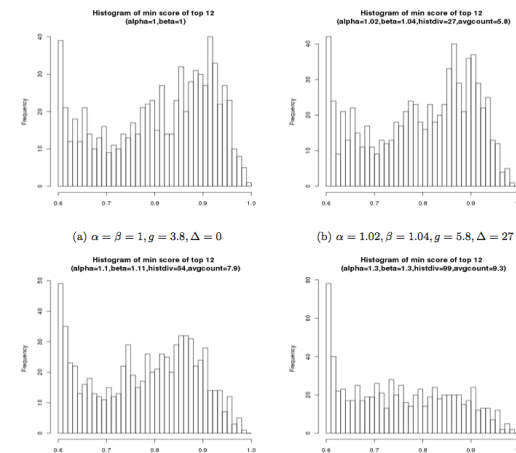
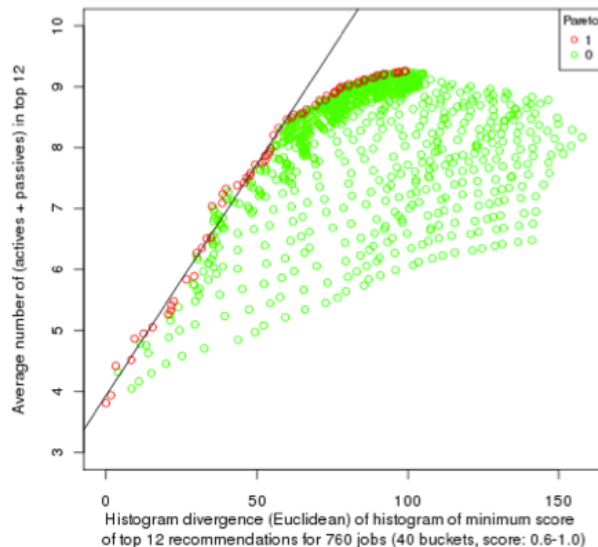
Multiple Objective Optimization

Constraint optimization problem

1. Rank top $K' > K$ results wrt to primary objective (e.g. relevance)
2. Perturb the ranking with a parametric function which leads to the inclusion of secondary objective(s) (e.g. engagement)
 - Measure the perturbation using a distance function wrt primary objective
 - Create a framework to quantify the tradeoff between the objectives

TalentMatch Use Case I

- Proxy for likelihood to answer hiring managers inquiries:
 - Job seekers are 16X more likely to answer than non-seekers
- Increase % of job seekers in in TalentMatch results
- Control for TalentMatch booking rate and sent emails rate



Distribution of min TalentMatch scores over 1 month of jobs posted on LinkedIn

TalentMatch Use Case II

Table 2: InMail response rate ratio of treatment group over control group. Data collected over a period of four weeks around May/27/2012-June/17/2012.

Treatment	Treatment/Control	95% CI
$\alpha = \beta = 1.15$	1.42	0.95 - 2.01
$\alpha = \beta = 1.07$	1.22	0.94 - 1.59

Table 3: Booking rate ratio of treatment group over control group. Data collected over a period of four weeks around May/27/2012-June/17/2012.

Treatment	Treatment/Control	95% CI
$\alpha = \beta = 1.15$	0.96	0.81 - 1.15
$\alpha = \beta = 1.07$	1.00	0.85 - 1.20

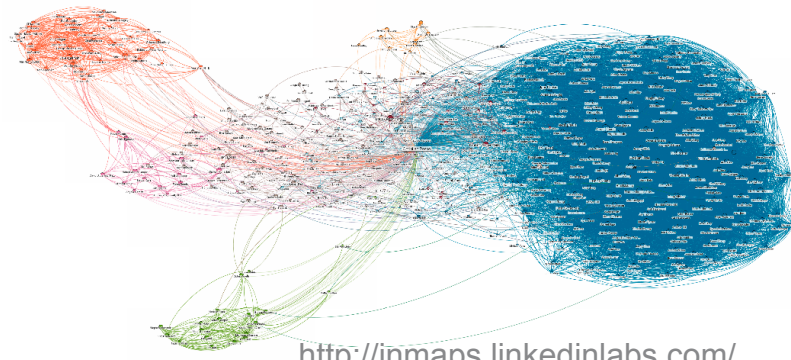
Table 4: Inmails/booking ratio of treatment group over control group. For perspective, the control group has approximately 2 InMails/booking. Data collected over a period of four weeks around May/27/2012-June/17/2012.

Treatment	Treatment/Control	95% CI
$\alpha = \beta = 1.15$	1.25	1.07 - 1.51
$\alpha = \beta = 1.07$	1.31	1.11 - 1.57

Data Sources for Feature Engineering

Social Graphs

Content



<http://inmaps.linkedinlabs.com/>

Behavior



Applied Researcher / Data Scientist: Data Analysis, Data Mining and Machine Learning
LinkedIn - Mountain View, CA (San Francisco Bay Area)

Job Description
The engineering culture at LinkedIn is based on building and integrating cutting-edge technologies while encouraging creativity, innovation and expansion. Our engineers consistently raise the bar for excellence, motivating each other to tackle challenges and take intelligent risks. The industry is moving fast and our engineers are right there with it! <http://engineering.linkedin.com/>

About Data Analysis, Data Mining and Machine Learning
We're looking for an Applied Researcher / Scientist to work on our massive semi-structured text and graph datasets. This role is ideal for a senior researcher who has both domain experience in data mining, information retrieval, or machine learning and a strong systems orientation. It offers the unique opportunity to conduct applied research and have a huge end-to-end impact on key LinkedIn product initiatives, including search relevance, ad targeting, information extraction, and recommendations.

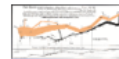
Desired Skills & Experience

- You are a computer science champ with strong coding ability and familiarity with Java, C++ or any other OOP language.
- You have worked with data mining toolkits like Weka, Mahout, R, NLTK, etc.
- You have experience with Hadoop or other MapReduce paradigms.
- You're familiar with information retrieval theories like Lucene / SOLR.
- You're interested in big data, in particular billions of semi-structured textual information, data mining, recommendations.
- You have pub. SDC, or P.
- You thrive in
- You enjoy us

Company Desc
Founded in 2003, LinkedIn is a successful Web 2.0 company. LinkedIn is a social network that connects professionals and businesses. It provides a platform for professionals to connect, share information, and advance their careers. LinkedIn is a leading provider of professional networking services, offering a wide range of features and tools for businesses and individuals alike.

Additional Info
Posted:
Type:
Experience:
Functions:
Industries:
Job ID:
Most Popular Discussions:
Here is a nice visualization of timeline events - an impressive interactive design.
Arab spring: an interactive timeline of Middle East protests
geopolitical event.
Ever since a riot in Tunisia turned itself to a riot in December 2010 in protest of the treatment of public, anti-democracy protesters have erupted across the Middle East. Our interactive timeline breaks key...
George Ingram 1 day ago · Jeffery, great presentation, we are now in 2012, is there a conclusion you can draw as to what will the Arab future desert?
LANKIN TODAY JAN 18
Defend Our Freedom to Share (or Why SOPA is a Bad Idea)
With the New Year, we have a new year of challenges ahead of us. One of the most important challenges is to defend our freedom to share. This is a critical issue for the future of the Internet, and it's one that we all have a stake in. We need to ensure that we can continue to share information, ideas, and creativity without being hindered by restrictive laws like SOPA. This is a fight that we need to win, and we need your help to do it.
Principal Data Scientist and Product Manager
LinkedIn
Public Company: 1001-5000 employees, LinkedIn
July 2008 - Present (3 years 7 months) | 36
Design and develop from ground up data infrastructure to the user experience, work
1. Highly scalable real-time content-base built on top of Hadoop, Lucene and Viole
- Personalized recommendations, Jobs
- Companies You May Want to Follow, Jobs For You Web Ads, Events You May Be Interested
in
- Similar recommendations for jobs, groups, companies, industries, profiles and events
- Matching solutions for our recruiting and marketing products: "Identify", "Relevant Engine", "Job Applicant Scoring", "Ad Targeting Engine"
- Crowd based recommendations: "Related Searches", "Viewers of This Profile Also Viewed", "People Who Viewed This Job Also Viewed", "Professional That Recommend This Product Also Recommended", "People Who Viewed This Company Also Viewed", "People Also Explored [these groups]"
In the press: <http://laturl.com/mmpaa>, <http://laturl.com/qc2>, <http://laturl.com/pby>.

Your Groups (48) Reorder



Visual Analytics



The R Project for Statistical Computing



Data Mining, Statistics, and Data Visualization

Data Mining, Statistics, and Data Visualization



Online Experimentation Professionals

Online Experimentation Professionals



Lesson 4: The Unreasonable Effectiveness of Big Data

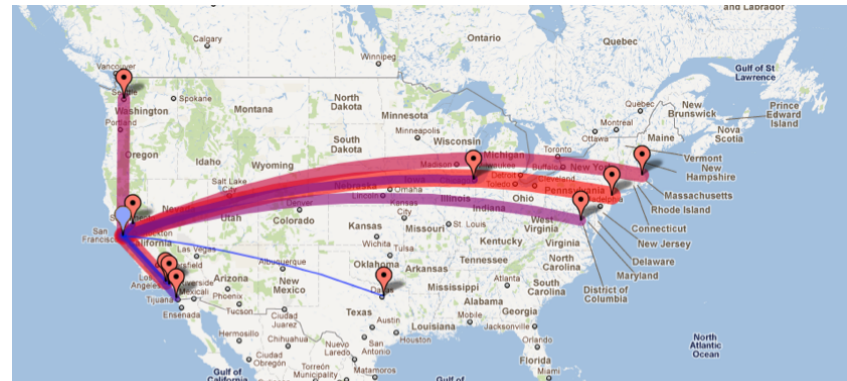
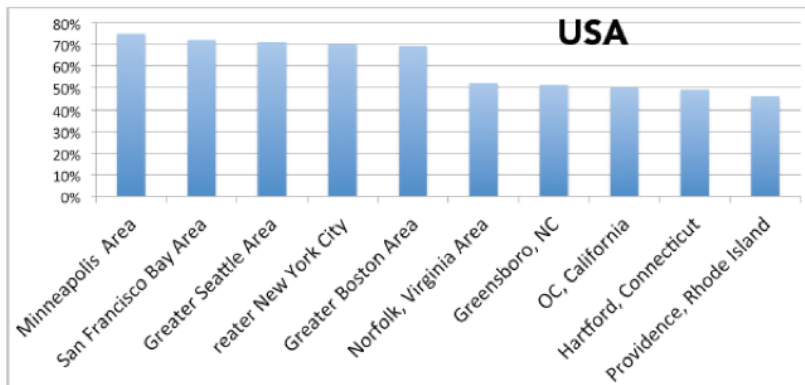
- Data jujitsu: slice and dice (smartly), then count
- Rich features engineered from profiles jujitsu
 - Job tenure distributions
 - Job transition probabilities
 - Related titles, industries, companies
 - Profile & job seniority
 - Impact on job recommendations:
 - 25% lift in views, viewers, applications, applicants
 - 90% drop in negative feedback

Lesson 4: The Unreasonable Effectiveness of Big Data (cont.)

- Region stickiness
- Related regions
 - Impact on job recommendations: 20% lift in views, viewers, applications, applicants
- Individuals propensity to migrate

Most vs Least sticky regions*

*what percentage of people stay in the same region when switching companies



Lesson 5: Data Labeling Can Be Daunting

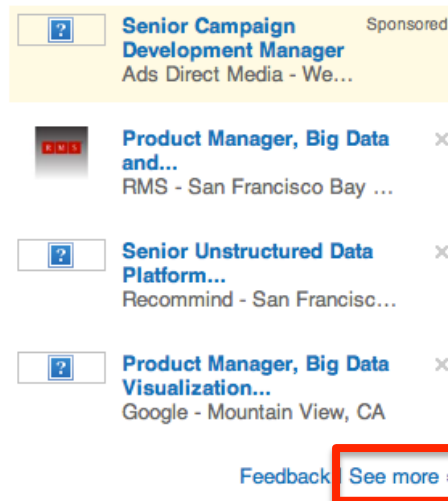
- **Historical data**
 - Similar objective
 - Unrelated processes, e.g., same session search selection
 - reduce presentation bias, position bias
 - What about intent bias?
- **Random suggestions**
 - Great with ads, company follows
 - Not for products with high cost of bad recommendations
 - jobs, alumni groups, ...
 - Not for similar recommendations
- **Crowdsourced manual labeling**
 - Very challenging
 - Pairwise comparison more suited than absolute rating → higher cost
 - Expert-sourced manual labeling

Lesson 6: Measure Everything

1. Implicit user feedback

- E.g. impressions, immediate actions (clicks), secondary actions
- Understand and optimize flows, e.g.,
 - Impact of 'See more' link and landing page
 - Conversion rate of a job view into a job application from various channels
- E.g. Homepage vs email vs mobile

JOBS YOU MAY BE INTERESTED IN



The screenshot shows a section titled "JOBS YOU MAY BE INTERESTED IN" with four job listings. Each listing includes a small icon, a job title, a company name, and a location. The first listing is a sponsored job for a "Senior Campaign Development Manager" at "Ads Direct Media - We...". The other three listings are for "Product Manager, Big Data and..." at "RMS - San Francisco Bay ...", "Senior Unstructured Data Platform..." at "Recommind - San Francisc...", and "Product Manager, Big Data Visualization..." at "Google - Mountain View, CA". At the bottom of the list, there are two links: "Feedback" and "See more »". The "See more »" link is highlighted with a red rectangular box.

Senior Campaign Development Manager Sponsored
Ads Direct Media - We...

Product Manager, Big Data and...
RMS - San Francisco Bay ...

Senior Unstructured Data Platform...
Recommind - San Francisc...

Product Manager, Big Data Visualization...
Google - Mountain View, CA

Feedback See more »

Lesson 6: Measure Everything (cont.)

2. Explicit user feedback

- Understand usefulness of the recommendation with unequal depth
 - Text based A/B test!
- Help prioritize future improvements
- Reveal unexpected complexities
 - E.g. 'Local' means different things for different members
- Prevent misinterpretation of implicit user feedback!

JOBS YOU MAY BE INTERESTED IN

 **Senior Campaign Development Manager** Sponsored
Ads Direct Media - We...

 **Product Manager, Big Data and...** ×
RMS - San Francisco Bay ...

 **Senior Unstructured Data Platform...** ×
Recommind - San Francisc...

 **Product Manager, Big Data Visualization...** ×
Google - Mountain View, CA

[Feedback](#) | [See more »](#)

From: Joseph Hicks-Malloy <jh041177@yahoo.com>
Date: August 12, 2012 5:31:32 PM PDT
Subject: [Site Feedback] Job Recommendations Feedback
To: LinkedIn Feedback <suggestions@linkedin.com>

The following comment was submitted on the LinkedIn website on August 13, 2012 12:31 AM by Joseph Hicks-Malloy:

Your job suggestions are very helpful keep up the great work

From: Steve Lane <Steve.Lane850@gmail.com>
Date: August 12, 2012 6:00:15 PM PDT
Subject: [Site Feedback] Job Recommendations Feedback
To: LinkedIn Feedback <suggestions@linkedin.com>

The following comment was submitted on the LinkedIn website on August 13, 2012 1:00 AM by Steve Lane:

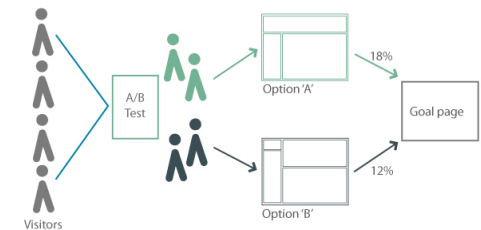
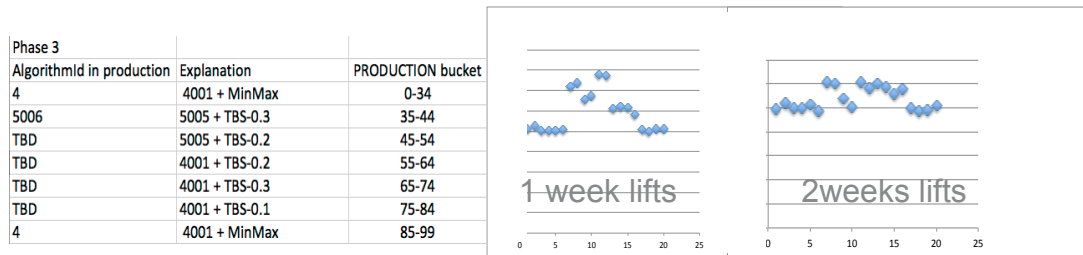
Need jobs in memphis tn.

Steve

Lesson 7: Beware of some A/B testing pitfalls

1. Novelty effect

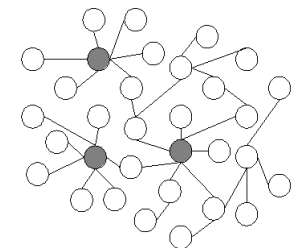
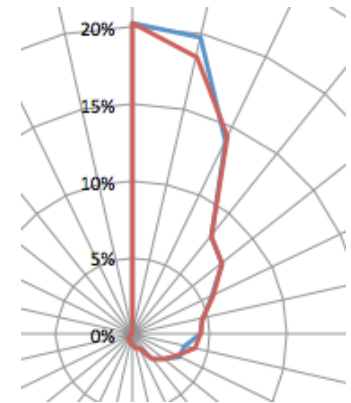
- E.g., new job recommendation algorithms have week-long novelty effect that shows lifts twice the stationary (real) one



2. Cannibalization

- Zero-sum game or real lift?

3. Random sampling destroys network effect



Lesson 8: One Unified Platform, One API

- **Scaling innovation**
 - Cross-leverage improvements between products
 - Shared knowledgebase
- **Maintainability**
 - Production serving and tracking
 - Infrastructure for complete upgrades
- **Performance**
 - Billions of sub second computations

Open Source Technologies



Zoie



Bobo



Kafka



Voldemort



Conclusion

- Social/professional networks are a new frontier for recommender systems
- Still many open questions:
 - How do we define and measure engagement?
 - What is the utility of a recommendation for the member?
 - What is the value of a recommendation for the network?
 - How do we reconcile utility and value when they conflict?
 - How do we network A/B test without tears?
 - ...
- Learn as you go
 - Track everything and invest in forensics analytics
 - Breadth – understand holistic impact
 - Depth – understand flows

References

- M. Rodriguez, C. Posse and E. Zhang. 2012. Multiple Objective Optimization in Recommendation Systems. To appear in *Proceedings of the Sixth ACM conference on Recommender systems (RecSys '12)*
- M. Amin, B. Yan, S. Sriram, A. Bhasin and C. Posse. 2012. Social Referral : Using Network Connections to Deliver Recommendations. To appear in *Proceedings of the Sixth ACM conference on Recommender systems (RecSys '12)*
- A. Reda, Y. Park, M. Tiwari, C. Posse and S. Shah. 2012. Metaphor: Related Search Recommendations on a Social Network. 2012. To appear in *Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM '12)*
- B. Yan, I. Bajaj and A. Bhasin. 2011. Entity Resolution Using Social Graphs for Business Applications. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM '11)*