Ahmed Metwally

(805) 403-9725

ametwally@gmail.com

Employment History (excluding full-time jobs and internships before 2006):

April 2008 - Now: Senior Software Engineer at Google, AdSpam (Ad Traffic Quality) Team.

Responsibilities:

1- Leadership:

a. Thought leadership: I proposed all the projects below, and was the main designer of the systems, the algorithms, the dependencies and data contracts between the system components, etc. These projects resulted in patents and publications based on the algorithms innovation and scalability, as well as the novelty in the techniques of fraud detection.

b. Project management: For all the projects below, I performed project management tasks including devising the project idea, holding brainstorming and regular meetings, scoping the project, breaking the project into tasks, estimating resources and time for completion, managing dependencies between engineers, monitoring progress, etc.

c. Cross-team collaboration: Multiple abuse-detection teams use the projects below. The projects are used to detect not only ad-click fraud, but also to detect abuse on search queries, YouTube likes and subscribes, local reviews, what's hot, comments, follows, re-shares, etc.

2- Major Projects:

a. Clustering Abusive Entities: The system is motivated by abusers impersonating several entities (e.g., opening several AdSense accounts) to distribute their activity among the entities, generate little fraud from each entity and stay under the radar level. The system evaluates the pair-wise similarity in a gigantic set of entities, and forms a similarity graph. The system then clusters the similarity graph to discover numerous communities of abusive entities. The communities (clusters) discovered vary drastically in properties (size, connectivity, isolatedness, etc.). The system allows for clustering on multiple corroborative signals by super-positioning clusters across signals. The system also provides a plugin for auto-tuning the different parameters (which similarity metric, minimum similarity, and the cluster properties for each dimension, which clustering algorithms, etc.) given the appropriate hooks for estimating the impact. (Publications: MF12, WMP13, MPDF15. Tools: MapReduce-C++, Python)

b. Estimating the Risk of Accounts: The system is motivated by the Google Ads subdivision willing to take different levels of liabilities based on how trustworthy the AdSense account is. The system maintains an estimate of how risky an AdSense account is using multiple sign-up, log-in, traffic, and

content signals. The system issues multiple scores based on the suspicious behavior expected from the account. (Tools: Python)

c. IP Size Estimation: The system analyzes the behavior of the users behind IPs, and estimates the number of users behind an IP (IP size), and predicting IPs' future sizes. The traffic is classified as anomalous based on the size of the IPs they come from, and is filtered accordingly. (Publications: MP11, SM12, MSPC14. Tools: MapReduce-C++, Python, R)

d. Histogram Filters of Traffic: The system represents several traffic signals of an entity as histograms, and filters the traffic from different bins based on the expected proportion of traffic in each bin. The system also allows for representing each entity (e.g., AdSense account) as multiple histograms, and classifying the entity as fraudulent or not based on how its histograms deviate from the expected. (Publications: SM12, MSPC14. Tools: Google-Sawzall, Python)

June 2007 - April 2008: Software Engineer at Oracle, Query Optimization Group.

Responsibilities:

1- Build a random generator of database workloads, including schemas, tables, and queries. I was responsible for the random query generation given a schema DDL. The objectives are as follows:

   a. The frequency of generating the various constructs of the DML language could be dynamically changed at run time. For each generated query, the framework examines which parts of Oracle's sever is being tested. The server subroutines (e.g. query transformations) that are not frequently tested using the generated queries should feedback into the probability knobs that generate the DML constructs. These knobs should automatically adapt and generate queries with constructs that test the non-tested server subroutines.

   b. The semantic context of the query should be discoverable by traversing the parse tree. Hence, any part of the query can be altered while not violating the semantic correctness of the query. The goal is to grow the complexities of the generated queries gradually for easier debugging of queries causing the server to crash or producing wrong results

October 2006 - March 2007: Internship at Ask.com.

Responsibilities:

1- Designing and implementing a new approximate algorithm for counting the number of distinct elements in a huge dataset using only one scan on the data and limited space. Doing the statistical analysis to prove the error guarantees. Implementing all the existing algorithms in the stream distinct counting literature to compare the experimental results on real data. (Publications: MAE08. Tools: C++)

2- Analyzing customer behavior, and designing a metric to measure customer retention.

3- Assisting Divyakant Agrawal (Vice President of Warehousing) in designing a warehouse of click streams.

Jun 2006 - Sep 2006: Internship at Google Inc., Infrastructure Division, Logs Analysis Team.

Responsibilities:

1- Designing and implementing a library for stream algorithms over variable-sized sliding windows. Designing and implementing algorithms to find approximate frequent elements and exact Min/Max-k elements in amortized constant time per stream element. (Tools: C++)

Education:

2002 - 2007: PhD; UC Santa Barbara, Department of Computer Science.

- Advisors: Prof. Amr El Abbadi and Prof. Divyakant Agrawal.
- Thesis title: "*On-Line Data Forensics for Fraud Detection in Internet Advertising*".

2002 - 2006: MS; UC Santa Barbara, Department of Computer Science.

2000 - 2002: MS Candidate; Alexandria University, Faculty of Engineering, Computer and Systems Engineering Department. Degree unfinished.

1995 - 2000: BS; Alexandria University, Faculty of Engineering, Computer and Systems Engineering Department.

Publications:

Journal Articles and Book Chapters:

- [M16] **Ahmed Metwally**: "*Frequent Items on Streams*". Springer Publishers' **Encyclopedia of Database Systems Ed. 2**, Tamer Özsu and Ling Liu (Eds.), to appear, 2015.
- [MSPC14] **Ahmed Metwally**, Fabio Soldo, Matt Paduano, Meenal Chhabra: "*Large-Scale Network Traffic Analysis for Estimating the Size of IP Addresses and Detecting Traffic Anomalies*". **Large Scale and Big Data**, Sherif Sakr, Mohamed Medhat Gaber (Eds.), pages: 435-462, 2014.
- [EMAE14] Fatih Emekci, **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Dividing secrets to secure data outsourcing*". Information Sciences, Vol. 263, pages: 198-210, 2014.
- [MF12] **Ahmed Metwally**, and Christos Faloutsos: "*V-SMART-Join: A Scalable MapReduce Framework for All-Pair Similarity Joins of Multisets and Vectors*". **PVLDB** Proceedings of the Very Large Data Bases Endowment, Vol. 5, No. 8, pages: 704-715, 2012.
- [AEEMW11] Divyakant Agrawal, and Amr El Abbadi, Fatih Emekci, **Ahmed Metwally**, and Shiyuan Wang: "*Secure Date Management Service on Cloud Computing Infrastructures*". Springer Publishers' **New Frontiers in Information and Software as Services - Service and Application Design Challenges in the Cloud**, Divyakat Agrawal, K. Selçuk Candan and Wen-Syan Li (Eds.), pages: 57-80, 2011.
- [M09] **Ahmed Metwally**: "*Frequent Items on Streams*". Springer Publishers' **Encyclopedia of Database Systems Ed. 1**, Tamer Özsu and Ling Liu (Eds.), pages: 1175-1179, 2009.

- [MEAE08] **Ahmed Metwally**, Fatih Emekci, Divyakant Agrawal, and Amr El Abbadi: "*SLEUTH: Single-pubLisher attack dEtection Using correlaTion Hunting*". **PVLDB** Proceedings of the Very Large Data Bases Endowment, Vol. 1, No. 2, pages: 1217-1228, 2008.
- [MAE06] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*An Integrated Efficient Solution for Computing Frequent and Top-k Elements in Data Streams*". ACM **TODS** Transactions On Database Systems, Vol. 31, No. 3, pages: 1095-1133, September 2006.

Conference Papers:

- [MPDF15] **Ahmed Metwally**, Jia-Yu Pan, Minh Doan, Christos Faloutsos *"Scalable Community Discovery from Multi-Faceted Graphs*". IEEE **BigData** Conference, 2015.
- [WMP13] Ye Wang, **Ahmed Metwally**, and Srinivasan Parthasarathy: *"Scalable all-pairs similarity search in metric spaces*". ACM **SIGKDD** International Conference on Knowledge Discovery and Data Mining, pages: 829-837, 2013.
- [SM12] Fabio Soldo, and **Ahmed Metwally**: "*Click Fraud Detection Based on Click Size Distribution*". IEEE **INFOCOM** International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, pages: 2005-2013, 2012.
- [MP11] **Ahmed Metwally,** and Matt Paduano: "*Estimating the Number of Users behind IP Addresses for Combating Abusive Traffic*". ACM **SIGKDD** International Conference on Knowledge Discovery and Data Mining, pages: 249-257, 2011.
- [KTPMDCB09] Carmelo Kintana, David Turner, Jia-Yu Pan, **Ahmed Metwally,** Neil Daswani, Erika Chin, and Andrew Bortz: "*The Goals and Challenges of Click Fraud Penetration Testing Systems*". IEEE **ISSRE** International Symposium on Software Reliability Engineering, 2009, paper number 187. Available at http://www.google.com/adwords/adtrafficquality/tech.html.
- [AEEM09] Divyakant Agrawal, Amr El Abbadi, Fatih Emekci, and **Ahmed Metwally**: "*Database Management as a Service: Challenges and Opportunities*". IEEE **ICDE** International Conference on Data Engineering, pages: 1709-1716, 2009.
- [MAE08] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Why Go Logarithmic if We Can Go Linear? Towards Effective Distinct Counting of Search Traffic*". **EDBT** International Conference on Extending Database Technology, pages: 618-629, 2008.
- [MAEZ07] **Ahmed Metwally**, Divyakant Agrawal, Amr El Abbadi, and Qi Zheng: "*On Hit Inflation Techniques and Detection in Streams of Web Advertising Networks*". IEEE **ICDCS** International Conference on Distributed Computing Systems, paper number 52, 2007.

- [BMAE07a] Nagender Bandi, **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Fast Algorithms for Data Streams using Associative Memories*", ACM **SIGMOD** International Conference on Management of Data, pages: 247-256, 2007.
- [BMAE07b] Nagender Bandi, **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*TCAM-conscious Algorithms for Data Streams*", IEEE **ICDE** International Conference on Data Engineering, pages: 1342-1344, 2007.
- [MAE07] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*DETECTIVES: DETEcting Coalition hiT Inflation attacks in adVertising nEtworks Streams*". The International **WWW** Conference, pages: 241-250, 2007.
- [MAE05a] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Using Association Rules for Fraud Detection in Web Advertising Networks*". **VLDB** International Conference on Very Large Data Bases, pages: 169-180, 2005.
- [MAE05b] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Duplicate Detection in Click Streams*". The International **WWW** Conference, pages: 12-21, 2005.
- [MAE05c] **Ahmed Metwally**, Divyakant Agrawal, and Amr El Abbadi: "*Efficient Computation of Frequent and Top-k Elements in Data Streams*". **ICDT** International Conference on Database Theory, pages: 398-412, 2005. **TEST OF TIME AWARD, 2015.**
- [FAEM04] Ying Feng, Divyakant Agrawal, Amr El Abbadi, and **Ahmed Metwally**: "*Range CUBE: Efficient Cube Computation by Exploiting Data Correlation*". IEEE **ICDE** International Conference on Data Engineering, pages: 658-670, 2004.

Ongoing Research:

- [PMPFS] Spiros Papadimitriou, **Ahmed Metwally**, Jia-Yu Pan, Christos Faloutsos, and Ramakrishnan Srikant: "*Practical Forecasting for Noisy Time Series*". Submitted for publication.
- [MAEE] **Ahmed Metwally**, Divyakant Agrawal, Amr El Abbadi, and Fatih Emekci: "*Why Go Logarithmic if We Can Go Linear? Reviving Linear Counting for Stream Applications*". Journal Manuscript under preparation.

Invited Talks (excluding presentations of conference papers and job talks):

- "*From Dups to Rings*". Google Publisher Summit, July 2014.
- "*SimCluster: Catching Fraudster Rings*". Google Anti-Abuse Summit, May 2016.
- "*SimCluster: Large Scale Graph Mining for Fraud Detection*". Google Research Conference, November 2014.
- "*SimCluster: the Ring Buster*". Google Anti-Abuse Summit, May 2014.

- "*Detecting Coordinated Click Fraud Attacks at Scale*". University of California, Santa Barbara, Department of Computer Science, April 2013.
- "*Can the Size of IP Addresses Help Click Fraud Detection*". Google company-wide tech-talk, March 2012.
- "*Click Fraud Detection Based on the Size of IP Addresses*". University of California, Santa Barbara, Department of Computer Science, February 2012.
- "*Large Scale Estimation and Forecasting in Practice*". The 3rd Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2011) in conjunction with SIGKDD 2011.

Academic Services:

Program Committee Member: SIGMOD 2010, AICCSA 2013 (2 tracks), SIGMOD 2015.

Dissertation Committee Member: Fabio Soldo, UCI, 2011.

Journal Reviewer: ACM TODS Transactions On Database Systems.
ACM TOSN Transactions On Sensor Networks.
CACM Communications of the ACM.
The VLDB Journal
IEEE Communications Letters
IEEE TC Transactions on Computers
IEEE TKDE Transactions on Knowledge and Data Engineering
Frontiers in ICT

External Reviewer: with Prof. D. Agrawal, Prof. A. El Abbadi, and Dr. Mohamed ElFeky, ACM-GIS, EDBT, ICDCS, ICDE, ICDT, PODS, SIGMOD, VLDB, WWW.

Mentorship: Mentored interns Amr Ebaid (Purdue, Summer 2009), Fabio Soldo (UCI, Summer and Fall 2010), Ye Wang (OSU, Summer 2011), Meenal Chhabra (RPI, Summer 2012), Ramya Korlakai Vinayak (CalTech, Summer 2014), and Luis Pineda (University of Massachusetts Amherst, Summer 2015) in the Ad Traffic Quality Team at Google.