STEVEN E. WHANG

1600 Amphitheatre Parkway,
Mountain View, CA 94043
Homepage: http://infolab.stanford.edu/∼euijong

Phone: (650) 796-6006
Email: euijong@gmail.com,
swhang@cs.stanford.edu

RESEARCH INTERESTS

Database-Artificial Intelligence (DB-AI) Integration, Big Data Analytics, Information Integration, Knowledge Systems, Machine Learning

EDUCATION

Stanford University, Stanford, CA — June 2012
Ph.D. in Computer Science
Advisor: Prof. Hector Garcia-Molina
Thesis: *Data Analytics: Integration and Privacy*

Stanford University, Stanford, CA — June 2007
M.S. in Computer Science
Distinction in Research (Database Specialization)
Advisor: Prof. Jennifer Widom
Thesis: *Generic Entity Resolution with Negative Rules*

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea — Feb. 2003
B.S. in Computer Science
GPA: 4.05/4.3 (Class 2003: 1^{st} among 96 in Computer Science Department, KAIST President's Prize)

PROFESSIONAL EXPERIENCE

- **Research Scientist, Google Research, Mountain View** Dec. 2012 – Present
  Managers: Dr. Alon Halevy, Dr. Neoklis Polyzotis
  Worked on understanding structured data on the Web. Technical leader for the Biperpedia project – an ontology for search applications. (Papers: VLDB '14, EMNLP '14, WebDB '15, WWW '16). Worked on a search engine for organizing Google's datasets. (Papers: SIGMOD '16, IEEE Data Bull. '16). Currently working on Big data management for large-scale machine learning systems. (Tutorial: SIGMOD '17, Paper: KDD '17)
- **Postdoctoral Researcher, Stanford University CS Department (InfoLab)** July 2012 – Dec. 2012
  Advisor: Prof. Hector Garcia-Molina
  Worked on crowdsourcing algorithms and interfaces for entity resolution.
- **Research Assistant, Stanford University CS Department (InfoLab)** June 2006 – June 2012
  Advisor: Prof. Hector Garcia-Molina
  Proposed general techniques for improving the accuracy, scalability, and functionality of entity resolution (also known as deduplication or record linkage). Also applied entity resolution to data privacy, where managing information leakage is becoming a critical problem.
- **Research Intern, Yahoo! Research (APEX group), Santa Clara** June 2008 – Dec. 2008
  Mentors: Dr. Jayavel Shanmugasundaram, Dr. Ramana Yerneni
  Developed efficient indexing techniques for complex boolean expressions using inverted lists. Applications include display advertising and, in general, publish/subscribe systems. Filed two patents with this work. (Paper: VLDB '09)
- **Research Intern, HP Labs (Advanced Business Intelligence Group), Palo Alto** Summer 2007
  Mentors: Dr. Malu Castellanos, Dr. Umeshwar Dayal
  Developed an upfront recommender for physical database design for the HP Neoview Data Warehouse product. (Paper: IEEE ICDE '09)

- **Software Engineer, Orom Info. Co. Ltd., Daejeon, Korea** Jan. 2003 – May 2005
  Developed and maintained a nation-wide digital library web service for elementary and high schools in Korea. (Substitute military service)

HONORS AND FELLOWSHIPS

- Best Paper Award, WebDB Workshop May 2015
- IBM PhD Fellowship 2011–2012
- School of Engineering Fellowship, Stanford University 2007–2008
- Korea Foundation for Advanced Studies (KFAS) Fellowship 2005–2007
- KAIST President's Prize for Academic Excellence at Commencement Feb. 2003

PUBLICATIONS

REFEREED CONFERENCE OR WORKSHOP PAPERS

1. Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, Chiu Yuen Koo, Lukasz Lew, Clemens Mewald, Akshay Naresh Modi, Neoklis Polyzotis, Sukriti Ramesh, Sudip Roy, **Steven Euijong Whang**, Martin Wicke, Jarek Wilkiewicz, Xin Zhang, Martin Zinkevich, "TFX: A TensorFlow-Based Production-Scale Machine Learning Platform," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1387-1395, Halifax, Nova Scotia, Canada 2017. (Co-corresponding author)
2. Mina Farid, Ihab Ilyas, **Steven Euijong Whang**, Cong Yu, "Lonlies : Estimating Property Values for Long Tail Entities," In *Proc.* 39^{th} *Annual Int'l ACM SIGIR Conf. on Research and Development on Information Retrieval*, pp. 1125–1128, Pisa, Italy, July 2016. (Demo)
3. Alon Halevy, Flip Korn, Natalya Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, **Steven Euijong Whang**, "Goods: Organizing Google's Datasets, " In *Proc. 2016 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD)*, pp. 795–806, San Francisco, June 2016. (Acceptance rate 19%)
4. Alon Halevy, Natalya Noy, Sunita Sarawagi, **Steven Euijong Whang**, Xiao Yu, "Discovering Structure in the Universe of Attribute Names," In *Proc.* 25^{th} *Int'l Conf. on World Wide Web (WWW)*, pp. 939–949, Montreal, Canada, Apr. 2016. (Co-first and corresponding author; Acceptance rate 15.8%)
5. Dana Movshovitz-Attias, **Steven Euijong Whang**, Natalya Noy, Alon Halevy, "Discovering Subsumption Relationships for Web-Based Ontologies," In *Proc.* 18^{th} *International Workshop on the Web and Databases (WebDB)*, pp. 62–69, Melbourne, Australia, May 2015. (**Best paper award**; Corresponding author)
6. Mohamed Yahya, **Steven Euijong Whang**, Rahul Gupta, Alon Halevy, "ReNoun: Fact Extraction for Nominal Attributes," In *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 325–335, Doha, Qatar, Oct. 2014. (Corresponding author)
7. Rahul Gupta, Alon Halevy, Xuezhi Wang, **Steven Euijong Whang**, Fei Wu, "Biperpedia: An Ontology for Search Applications," In *Proc.* 40^{th} *Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 505–516, Hangzhou, China, Sept. 2014. (Proceedings of the VLDB Endowment, Volume 7 Issue 7; Co-first and corresponding author; Acceptance rate approx. 19.4%)
8. **Steven Euijong Whang**, Hector Garcia-Molina, "Disinformation Techniques for Entity Resolution," In *Proc.* 22^{nd} *ACM Int'l Conf. on Information and Knowledge Management (CIKM)*, pp. 715–720, San Francisco, California, Oct. 2013.
9. **Steven Euijong Whang**, Peter Lofgren, Hector Garcia-Molina, "Question Selection for Crowd Entity Resolution," In *Proc.* 39^{th} *Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 349-360, Trento, Italy, Aug. 2013. (Proceedings of the VLDB Endowment, Volume 6 Issue 6; Acceptance rate 22.7%)
10. **Steven Euijong Whang**, Hector Garcia-Molina, "A Model for Quantifying Information Leakage," In *Proc.* 9^{th} *VLDB Workshop on Secure Data Management (SDM)*, pp. 25–44, Aug. 2012.
11. **Steven Euijong Whang**, Hector Garcia-Molina, "Joint Entity Resolution," In *Proc.* 28^{th} *IEEE Int'l Conf. on Data Engineering (ICDE)*, pp. 294–305, Washington, DC, Apr. 2012. (Acceptance rate approx. 20%)
12. **Steven Euijong Whang**, Hector Garcia-Molina, "Managing Information Leakage," In *Proc.* 5^{th} *Biennial Conf. on Innovative Data Systems Research (CIDR)*, pp. 79–84, Pacific Grove, California, Jan. 2011.

13. **Steven Euijong Whang**, Hector Garcia-Molina, "Entity Resolution with Evolving Rules," In *Proc. 36th Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 1326–1337, Singapore, Sept. 2010. (Proceedings of the VLDB Endowment, Volume 3 Issue 1-2; Acceptance rate 15.8%)
14. David Menestrina, **Steven Euijong Whang**, Hector Garcia-Molina, "Evaluating Entity Resolution Results," In *Proc.* 36th *Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 208–219, Singapore, Sept. 2010. (Proceedings of the VLDB Endowment, Volume 3 Issue 1-2; Corresponding author; Acceptance rate 15.8%)
15. **Steven Euijong Whang**, Chad Brower, Jayavel Shanmugasundaram, Sergei Vassilvitskii, Erik Vee, Ramana Yerneni, Hector Garcia-Molina, "Indexing Boolean Expressions," In *Proc. 35th Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 37–48, Lyon, France, Aug. 2009. (Proceedings of the VLDB Endowment, Volume 2 Issue 1; Acceptance rate 17.9%)
16. **Steven Euijong Whang**, David Menestrina, Georgia Koutrika, Martin Theobald, Hector Garcia-Molina, "Entity Resolution with Iterative Blocking," In *Proc. 2009 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD)*, pp. 219–232, Providence, Rhode Island, June 2009. (Acceptance rate 15.9%)
17. Malu Castellanos, Ivo Jimenez, Neal Coddington, Hans Zeller, **Steven Whang**, Umeshwar Dayal , "QuickStart: An Upfront Client-based Design Advisor for Parallel Data Warehouses," In *Proc. 25th IEEE Int'l Conf. on Data Engineering (ICDE)*, pp. 1543–1546, Shanghai, China, Mar. 2009. (Demo)
18. Ki-Hoon Lee, Seo-Young Kim, **Euijong Whang**, Jae-Gil Lee, "A Practitioner's Approach to Normalizing XQuery Expressions," In *Proc. 11th Int'l Symposium on Database Systems for Advanced Applications (DASFAA)*, pp. 437–453, Singapore, Apr. 2006.

TUTORIALS

1. Neoklis Polyzotis, Sudip Roy, **Steven Euijong Whang**, Martin Zinkevich, "Data Management Challenges in Production Machine Learning," In *Proc. 2017 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD)*, pp. 1723-1726, Chicago, Illinois, May 2017. (Co-first author)

REFEREED JOURNAL PAPERS

1. **Steven Euijong Whang**, Hector Garcia-Molina, "Incremental Entity Resolution on Rules and Data," *VLDB Journal*, vol. 23, no. 1, pp. 77–102, 2014. (SCI Core, IF: 1.568)
2. **Steven Euijong Whang**, Hector Garcia-Molina, "Joint Entity Resolution on Multiple Datasets," *VLDB Journal*, vol. 22, no. 6, pp. 773–795, 2013. (SCI Core, IF: 1.701)
3. **Steven Euijong Whang**, David Marmaros, Hector Garcia-Molina, "Pay-As-You-Go Entity Resolution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1111–1124, 2013. (SCI Core, IF: 1.851)
4. **Steven Euijong Whang**, Omar Benjelloun, Hector Garcia-Molina, "Generic Entity Resolution with Negative Rules," *VLDB Journal*, vol. 18, no. 6, pp. 1261–1277, Feb. 2009. (SCI Core, IF: 4.517)
5. Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, **Steven Euijong Whang**, Jennifer Widom, "Swoosh: A Generic Approach to Entity Resolution," *VLDB Journal*, vol. 18, no. 1, pp. 255–276, Jan. 2009. (Co-first and corresponding author. Algorithms published in Chapter 21.7 of the textbook *Database Systems: The Complete Book*, 2nd Ed.) (SCI Core, IF: 4.517)

INVITED PAPERS

1. Alon Halevy, Flip Korn, Natalya Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, **Steven Euijong Whang**, "Managing Google's data lake: an overview of the Goods system," *IEEE Data Engineering Bulletin*, vol. 39, no. 3, pp. 5–14, Sept. 2016.
2. **Steven Euijong Whang**, Hector Garcia-Molina, "Developments in Generic Entity Resolution," *IEEE Data Engineering Bulletin*, vol. 34, no. 3, pp. 51–59, Sept. 2011.

TEACHING AND MENTORING

- **Faculty Research Award Host, Google Research**
  - Ihab Ilyas (Professor at University of Waterloo) 2014–2015
    Quality-Driven Targeted Information Extraction (Paper: SIGIR '16)
- **Intern Host, Google Research**

  - Yeounoh Chung (CS PhD student at Brown University) Spring 2017
    Data Slicing for Model Understanding
  - Sainyam Galhotra (CS Master student at Univ. Massachusetts) Summer 2016
    Service Recommendations for Datasets
  - Evangelos Papalexakis (CS PhD student at Carnegie Mellon University) Summer 2015
    Internationalizing Biperpedia
  - Dana Movshovitz-Attias (CS PhD student at Carnegie Mellon University) Summer 2014
    Discovering Subsumption Relationships for Web-Based Ontologies (Paper: WebDB '15, **best paper award**)

- **Intern Co-host, Google Research**
  - Mohamed Yahya (CS PhD student at Max Planck Institute) Winter 2014
    ReNoun: Fact Extraction for Nominal Attributes (Paper: EMNLP '14)
  - Manas Joglekar (CS PhD student at Stanford University) Summer 2013
    Finding Inter-Attribute Relationships in Biperpedia
  - Xuezhi Wang (CS PhD student at Carnegie Mellon University) Summer 2013
    Biperpedia: An Ontology for Search Applications (Paper: PVLDB '14)

- **Student Supervisor, Stanford University**
  - Peter Lofgren (CS PhD student at Stanford University) Winter 2012
    Question Selection for Crowd Entity Resolution (Paper: PVLDB '13)
  - David Marmaros (Stanford MSCS student from Google) Winter 2010
    Pay-As-You-Go Entity Resolution (Paper: IEEE TKDE '13)

- **Instructor, Stanford University**
  - CS245, Database System Principles Summer 2009
    Number of students: 34, instructor mean score: 4.24/5, course mean score: 4.29/5

- **Course Assistant, Stanford University**
  - CS245, Database System Principles (Taught by Prof. Hector Garcia-Molina) Winter 2009

INVITED TALKS

- "Data Analytics: Integration, Privacy, and Knowledge"
  - Distinguished Lecture Series, APWeb, Changsha, China Sept. 2014
  - Seoul National University, Seoul, Korea Nov. 2013
  - KAIST, Daejeon, Korea Nov. 2013
- "Data Analytics: Integration and Privacy"
  - Cornell University, Ithaca, NY May 2012
  - MIT, Cambridge, MA Apr. 2012
  - Google Research, Mountain View, CA Apr. 2012
  - Microsoft Research, Redmond, WA Apr. 2012
  - IBM Research Almaden, San Jose, CA Mar. 2012
  - IBM Research Watson, Hawthorne, NY Dec. 2011
- "Managing Information Leakage"
  - Technicolor, Palo Alto, CA Aug. 2011
  - *TRUST* seminar series, U.C. Berkeley, Berkeley, CA Apr. 2011
  - *TRUST* conference, Stanford University, Stanford, CA Nov. 2010
- "Stanford Entity Resolution Framework"
  - Microsoft Research, Mountain View, CA Dec. 2010

  - HP Labs, Palo Alto, CA Mar. 2010
  - Stanford-Berkeley Exchange, U.C. Berkeley, Berkeley, CA Apr. 2009
- "Entity-Resolution: Beyond the Basics"
  - InfoLab Workshop, Stanford University, Stanford, CA Apr. 2010
- "Swoosh: A Generic Approach to Entity Resolution"
  - Guest lecture given in the Stanford Course CS345C: Data Integration taught by Dr. Alon Halevy May 2008
- "Generic Entity Resolution with Negative Rules"
  - InfoLab/Hitachi Workshop, Stanford University, Stanford, CA Mar. 2008

SERVICE

- Program Committee: *PVLDB* (2018, 2016, 2015), *ACM SIGMOD* (2018, 2015, 2014, 2013), *IEEE ICDE* (2017, 2015, 2014), *EDBT* (2016)
- External Reviewer: *ACM SIGMOD* (2017), *PVLDB* (2014, 2013), *ACM CIKM* (2013), *WWW* (2013), *ACM SIGMOD* (2011), *ACM KDD* (2014, 2010)
- Journal Reviewer: *VLDB Journal*, *IEEE TKDE*, *ACM TODS*
- Publicity Co-Chair: PAKDD 2017
- Web Co-Chair: ICDE 2015
- Editor of the Stanford CS Redbook (introductory booklet for the CS Ph.D. program) 2011
- Faculty Search Committee Student Member, Computer Science Department, Stanford University 2010
- KAIST Alumni President at Stanford 2007–8

PATENTS

- Sergei Vassilvitskii, Ramana Yerneni, Jayavel Shanmugasundaram, Erik Vee, Chad Brower, **Steven Whang**, "System and Method for Automatic Matching of Highest Scoring Contracts to Impression Opportunities Using Complex Predicates and an Inverted Index," US20110016109-A1, filed July 2009.
- Sergei Vassilvitskii, Ramana Yerneni, Jayavel Shanmugasundaram, Erik Vee, Chad Brower, **Steven Whang**, "System and Method for Automatic Matching of Contracts to Impression Opportunities Using Complex Predicates and an Inverted Index," US20100262607-A1, filed Apr. 2009.

GENERAL INFORMATION

- Citizenship: Republic of Korea, United States

REFERENCES

Available upon request.