ROHIT KUMAR

(+34) · 6021 · 97987 ⋄ rohit4phy@gmail.com
http://www.essi.upc.edu/∼rkumar

RESEARCH AREA

Big Data, Data Stream Mining, Graph Stream Mining and Distributed Processing, Data Security and Privacy.

EDUCATION

Université Libre de Bruxelles (ULB), Belgium & Universitat Politécnica de Catalunya (UPC), Spain August 2014 - Now

- Joint PhD in Computer Science
- Advisor: Toon Calders & Alberto Abello
- Research Topic: Adaptive and Converging Graph Data Stream Mining

Chennai Mathematical Institute, India April 2008 - April 2011

- M.Sc in Computer Science
- Advisor: Madhavan Mukund
- Thesis Topic: Security aspects of online Assessment platform-Touchstone
- CGPA : 8.4

Hindu College Delhi University, India July 2004 - July 2007

- B.Sc(Hons) in Physics
- Percentage : 76

RESEARCH EXPERIENCE

TCS - Tata Research Development and Design Center April 2011 - September 2011
Researcher *Pune, India*

- Worked on data privacy and data security.

JNCSAR, IISC, Bangalore 2004 - 2007
Summer Intern *Bangalore, India*

- Did two summer training in Nanoscience and research in general.
- Developed and experimented with new techniques to generate Zinc oxide Nanotubes.

TEACHING EXPERIENCE

TCS - iGnite Feb 2008 - July 2010
Teaching Assistant *Chennai, India*

- Teaching Assistant for Java and Fundamental of Programming in TCS iGnite Traning Center.
- Exam co-coordinator for online Mooshak exams.

ULB, Brussels 2015 - 2016
Teaching Assistant *Brussels, Belgium*

- Teaching Assistant for Data ware house labs for IT4BI masters course.

PROFESSIONAL EXPERIENCE

Royal bank of Scotland June 2014 - July 2014
Software Designer *New Delhi, India*

· Worked as a software developer, developing new systems on Oracle Coherence.

Tata Consultancy Services Nov 2007 - May 2014
Assistant Consultant *India*

· Technical team lead for the development of an **Online Assessment** software.
· Managed a team of 30 software developers.
· Developed and deployed the complete security Eco-system for the online assessment platform.
· Designed new offerings and solutions for Education Sector.
· Taught **Algorithms** and **Java** to new employees as a part of learning and development program.

PUBLICATIONS

- Location Influence in Location-based Social Networks. (Accepted), WSDM 2017
- Information Propagation in Interaction Networks. (Conditional accepted), EDBT 2017
- Maintaining Sliding-Window Neighborhood Profiles in Interaction Networks. (Published) ECML/PKDD 2015

PATENTS

- System and method for automated competency assessment (US8915744) - Granted
- Secured Computer Based Assessment (US20140186812A1) - Applied
- Customized question paper generation (US20130084554) - Applied

TECHNICAL SKILLS

Programming	J2EE, Java, Scala, C++, python, JDBC, HTML, JavaScript
Databases	MySQL, Microsoft SQL, MongoDB, Apache Cassandra, HBase, Oracle Coherence
Frameworks	Hadoop, Apache Spark, Apache Storm, Kafka, Struts 2, Hibernate
Tools	SVN, CVS, github, Eclipse, RStudio, Octave, Matlab, gnuplot

OTHER ACHIEVEMENTS

- FNRS Scholarship for Doctoral Research.
- Best Technical Lead Award in ACTion2013 in SMB iON, TCS.
- Innovation of the year award southern region in TATA Innovista 2011.
- Senior Diploma in Arts and Painting from Ankan Kala Bibhag, Ravindra Bharti University in 2006.

Research Description and Motivation

Graph data is one of the most used data structures in computer science. Massive graphs are getting generated regularly by applications which maintain a relationship between different data entities, for example: the social network in social media, publications, and authors in DBLP, web pages and hyperlinks, sensor data, etc. Analyzing these graphs to compute the graph properties like connectivity, shortest path, node distance, etc. is a well know problem. However, traditional methods require multiple passes over the huge graphs. Hence analyzing them under the streaming model is becoming more and more popular. Unlike static graph mining techniques for large graphs, graph stream mining possesses a more complex challenge where the large graph is constantly updating and evolving over time by the addition of new edges and the continuous deletion of old edges. Calculating properties such as the ones mentioned above on the evolving graph is a very interesting research problem in graph data stream mining. For example finding the shortest path between all pairs of nodes in a graph is an interesting problem for large graphs like social media as it could be used to solve numerous graph analysis problems such as centrality computation, community detection and node separation. One of the approaches is to create synopses of the graphs to preserve these properties. These synopses are sparse sub graphs of the original graph which could be used to approximate the properties of the original graph. The size of these synopses is very small compared to the original graph and hence can be kept in memory for faster mining. Another approach to handle massive graphs are by using distributed and parallel systems like Apache Spark GraphX, Graph Engine, Apache Giraph etc. to partition the graph data over different systems and manage the memory usage and computation time. These technique works on the principle of divide and conquers by dividing the computation and data over different machines. For dynamic graph data, however, doing a proper partitioning of data over different machines, to avoid too many communication, is much more complicated than for a data-set consisting of independent records or a static graph.

In my Ph.D. research study, we are using a combination of both approaches to solve some of the graph data stream mining problems. We first developed a novel incremental algorithm to maintain "Neighbourhood Profile" sketch of all the nodes in a graph using an extension of HyperLogLog sketch. Then we developed some variation of this sketch and used it for studying information flow in interaction network like micro-blog based social network twitter and also on location based social networks like FourSquare. Currently, I am working in developing a distributed version of our proposed algorithms in SPARK-GraphX and trying to address problems related to streaming graph and graph partitioning strategy to improve performance.