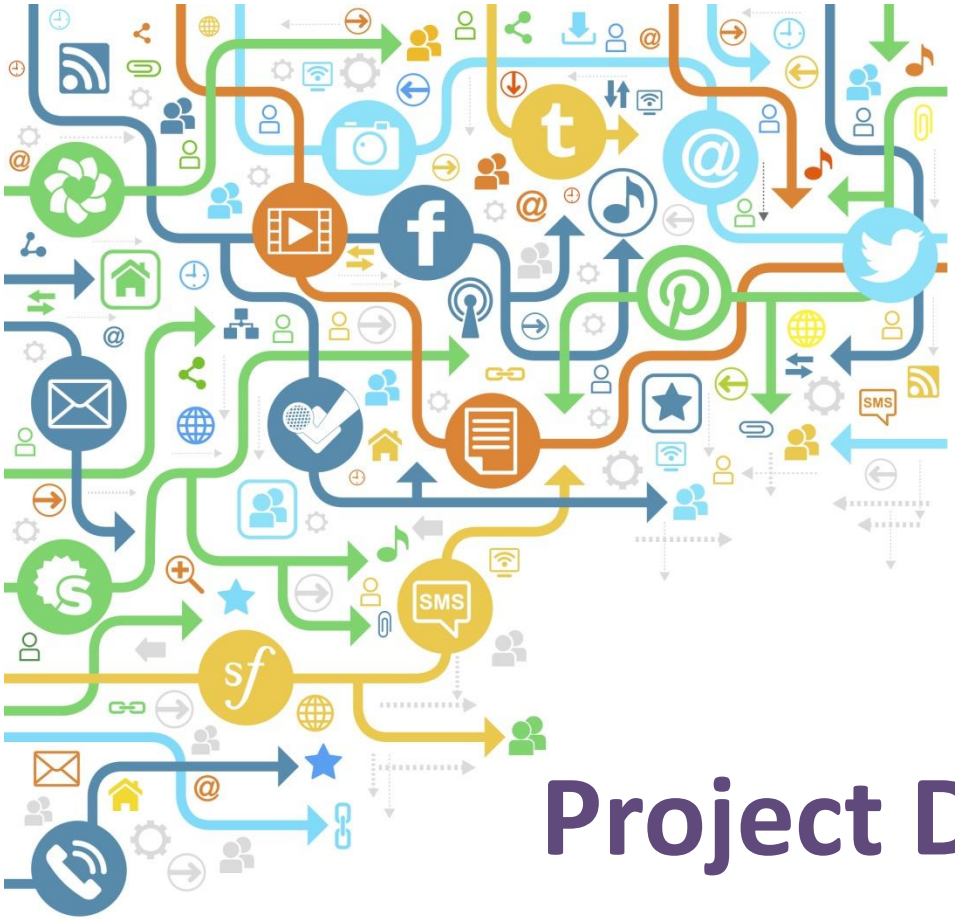

Project Discussion

Barna Saha

AT&T Lab-Research

What is required !

- Ideally ~10 projects
  - Make a group of 3 with matching interest
  - Else share your interest in the blog and I will create the groups
- How to select project
  - Gloss over bunch of papers on topics that interest you
    - THINK
      - Is there anything to improve upon ?
      - Can there be an interesting variant ?
      - Can we solve this problem on a different model ?
      - Can there be an interesting application where the work can be applied ? What changes are required ?
- START EARLY

Timeline

- October 3rd: Name of 6 papers (2 each)
  - Allocate time with me during office hours to describe the papers
  - Each group should allocate one hour time in total
  - All groups must meet me before November 3rd.
- October 17th: Two page project proposal (all submissions in pdf format written in latex)
  - Discuss what you want to work on your project much before this deadline, so that I don't have to disapprove any of your project.
- November 14th: One page progress report
- December 5th: Project write-up due
- 20-30 minutes in class presentation (Nov 21st, Dec 5th)
- DEADLINES ARE FIRM—LATE SUBMISSIONS WILL BE HIGHLY PENALIZED.

Scribe

- 21 lectures, 29 registered students !
  - 4 problem sets
  - 2 students each will write the solution
    - Solution will be discussed in class/ office hour (depending on course progress)
    - Students should try to solve the problem set and participate in class discussion on the problem sets.
- SCRIBE DUE BY FOLLOWING MONDAY MIDNIGHT
  - LATE SUBMISSION
    - Tuesday-Before the class: 50% penalized
    - Afterwards: 100% penalized

Project Type

- Improve Bounds of a problem
  - Example:
    - An Optimal Algorithm for the Distinct Elements Problem, Kane, Nelson, Woodruff, PODS'10
      - Optimal space and update time
  - Improvement
    - Space
    - Time
    - Number of Rounds/ Passes
    - Approximation Factor
      - May be different norm

Project Type

- Find Interesting Variant, Develop Algorithms and Analyze
  - Example “Finding Interesting Correlations with Conditional Heavy Hitters”, Mirylenka, Palpanas, Cormode, Srivastava, ICDE’12
    - Conditional Heavy Hitter: conditionally frequent: when a particular item is frequent within the context of its parent item
      - Find popular destination under source—”locally popular”
  - Think
    - Are there other estimates where conditioning may help ?
    - Distributed processing ?

Project Type

- Consider Different Models, Design Algorithm, Analysis
  - Traditional→ streaming, map reduce
  - Can we take the help of crowd ?
    - Can this problem formulation be useful for image/text analysis ?

Project Type

- Find Interesting Application, Design Algorithms, Analyze
  - Example: Real Time Story Identification (best paper on vldb 2012)
- Stories can be identified via groups of tightly-coupled real-world entities, namely the people, locations, products, etc., that are involved in the story

- Is dense subgraph sufficient ?
- What guarantees are provided ?
- Does other models make sense ?
- How does it compare with other papers?

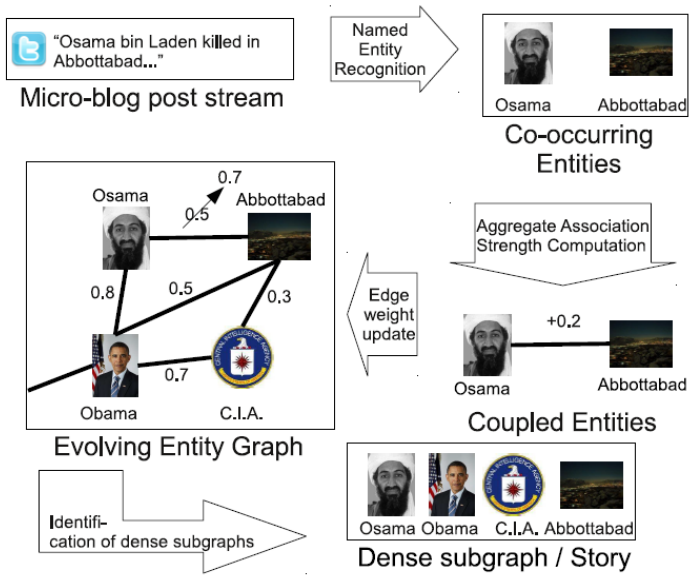

Figure 1: Real-time identification of "bin Laden raid" story,

When to do experiments ?

- If analysis alone is not sufficient to be published in a theory conference
- If target is a database or data mining or applied machine learning conference
  - Coding can be done in any programming language
  - All codes need to be submitted and need to be tested substantially
  - Same deadline as Dec 5th
- If the work is publishable in good theory conference (who decides ? Well! I do.)
  - NO experiment needed

Project Type

- Experimental Project
  - CANNOT be simple implementation of bunch of algorithms
    - NEED to have an objective
    - Example: Compare different heavy-hitter algorithms
      - Can you answer when is an algorithm good ? What characteristics of data does it depend upon ?
      - Which parameters are crucial ? How to tune those parameters ?
      - Is it possible to give a unifying framework that fits all ?
  - G. Cormode and D. Firmani. On unifying the space of l_0-sampling algorithms (read it to get an idea),
  - Heavy hitters, quantiles etc. have bee studied
  - Can you try graph based sketches or sketching for random matrices ?

Suggestions

- START EARLY
- READ PAPERS
- DIVIDE load among group members
- DISCUSS with me throughout—Friday afternoons are for you to discuss projects with me.