Steering User Behavior With Badges

Dan Huttenlocher
Cornell Computer Science Department
and Cornell NYC Tech

Joint work with Ashton Anderson, Jon Kleinberg, Jure Leskovec

Cornell University

Badges

- Have long history
  - Military medals, scouting, primary school, loyalty programs
- Not explicitly competitive
  - Contrast with ranking-based prizes
- Increasingly common online
  - Encouraging participation/contribution
  - Recognizing skills/achievement
  - Some backlash, ‘badge measles’

CORNELL NYCTECH

Social Psychology of Badges

- Broad range of possible individual value to earning badges [Antin and Churchill, 2011]
  - Goal setting
  - Instruction
  - Reputation
  - Status/Affirmation
  - Group identification

Badges and User Behavior

- Do badges affect user engagement and can we better characterize how?
  - Overall level of participation
  - Engagement in certain types of activities
- Can we help site designers define useful badges and systems if badges?
- In this work
  - Develop a model
  - Investigate in context of StackOverflow
  - Consider badge placement problem

CORNELL NYCTECH

stackoverflow Questions Tags Users Badges Unanswered

Connected components in a graph with 100 million nodes

Move apps to the cloud
without rewriting code.
Once you get it, you'll get it.

Windows Azure
FREE 90-DAY TRIAL

1

I am trying to get the list of connected components in a graph with 100 million nodes. For smaller graphs, I usually use the connected_components function of the Networkx module in Python which does exactly that. However, loading a graph with 100 million nodes (and their edges) into memory with this module would require ca. 110GB of memory, which I don't have. An alternative would be to use a graph database which has a connected components function but I haven't found any in Python. It would seem that Dex (API: Java, .NET, C++) has this functionality but I'm not 100% sure. Ideally I'm looking for a solution in Python. Many thanks.

python graph

share | improve this question

asked Jun 13 '12 at 13:48
user1453508
27 • 4

1 Answer

active oldest votes

1

SciPy has a connected components algorithm. It expects as input the adjacency matrix of your graph in one of its sparse matrix formats and handles both the directed and undirected cases.

Building a sparse adjacency matrix from a sequence of `(i, j)` pairs `adj_list` where `i` and `j` are (zero-based) indices of nodes can be done with

CORNELL
NYCTECH

Assumptions and Goals of Model

- Assume badges have value to users
- Assume each user has a preferred mix of actions, with a cost to deviate from that mix
- A user trades off between the preferred mix and the goal of winning the badge
- Want to understand effects of badges on overall engagement level and in “steering” user actions

CORNELL
NYCTECH

The Model

- A population of users and a site designer, with a fixed set of user actions

Ask Q

Answer Q

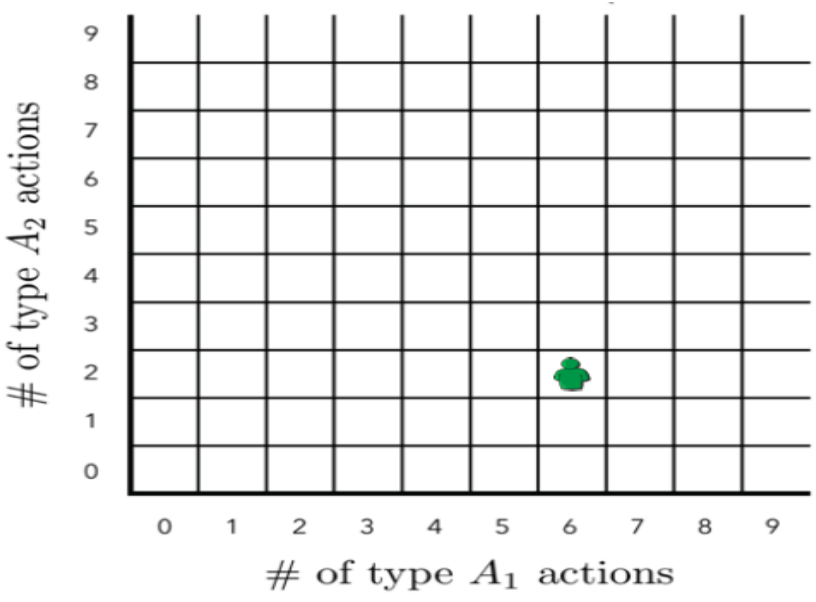
Vote on Q

⋮

CORNELL NYCTECH

The Action Space

- Action types A_1, A_2, … form space denoting number of actions of each type



CORNELL NYCTECH

User Model

- User has a preferred distribution, **p**, over action types
- At each time step user picks a probability distribution, **p**', and samples an action from it
- User incurs utility penalty for deviating from preferred distribution

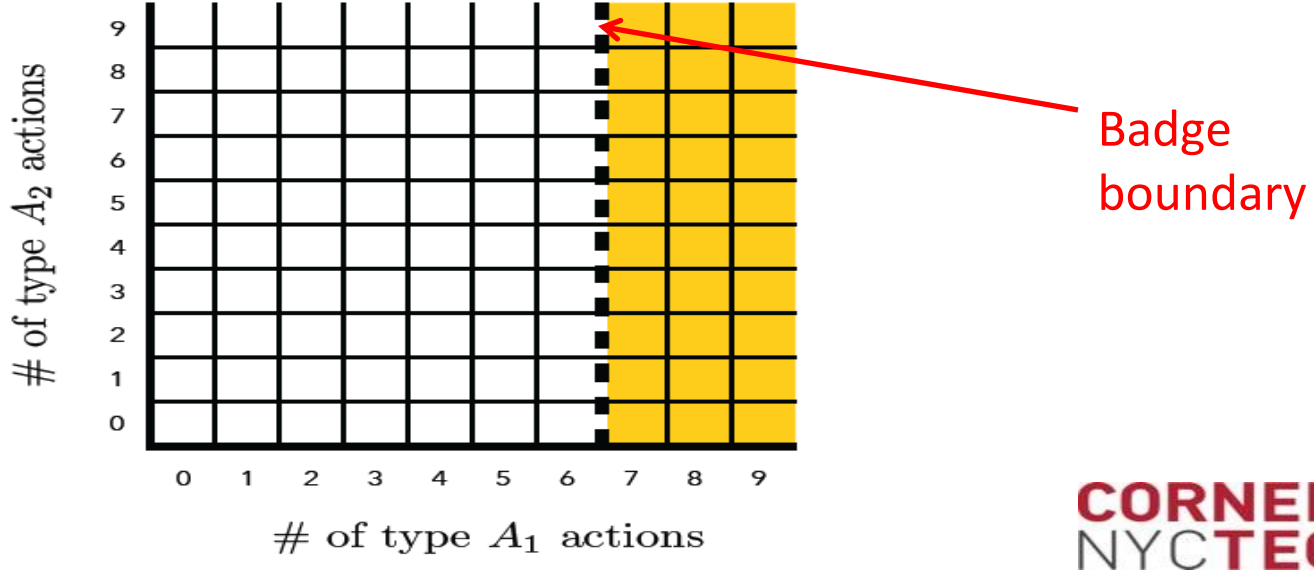
$$g(\mathbf{p}, \mathbf{p'}) = \|\mathbf{p} - \mathbf{p'}\|_2^2$$

- User survives to next step with probability Θ (generally Θ=.99)

CORNELL NYCTECH

Badges

- Set of badges, B, each b is subset of cells in action space and has value (utility) V_b



CORNELL NYCTECH

Utility Function

- User's utility composed of 3 parts
  - Value from badges won
  - Cost of deviating from p

Utility from won badges

Cost

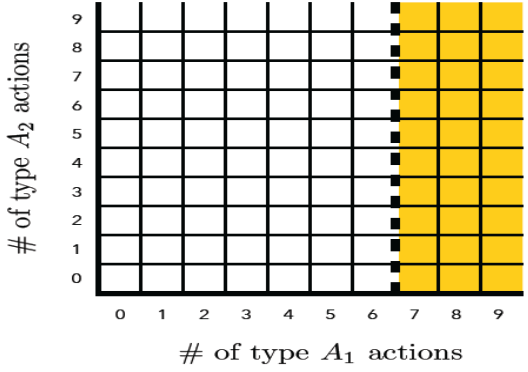
$$f(\mathbf{a}) = \overbrace{\sum_{b \text{ won}} V_b}^{\text{Utility from won badges}} + \underbrace{\theta[\mathbf{p}_\mathbf{a}^1 \cdot f(a_1+1, a_2) + \mathbf{p}_\mathbf{a}^2 \cdot f(a_1, a_2+1)]}_{\text{Expected utility of next state}} - \overbrace{g(\mathbf{p}, \mathbf{p}_\mathbf{a})}^{\text{Cost}}$$

Expected utility of next state

CORNELL NYCTECH

Optimization: One Targeted Dimension

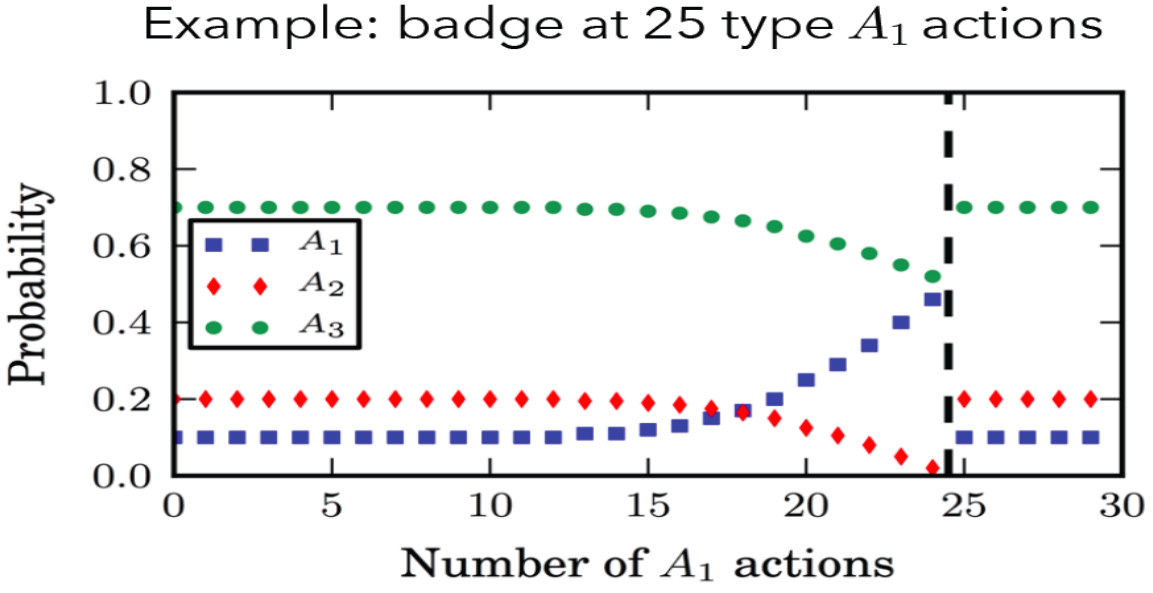
- Use dynamic programming to solve
- No reward for deviating from **p** past boundary
- Before boundary select $\mathbf{p}_a$ to maximize expected utility
- Collapses along A2 dimension
  - 1D problem, solve from boundary back to origin



CORNELL NYCTECH

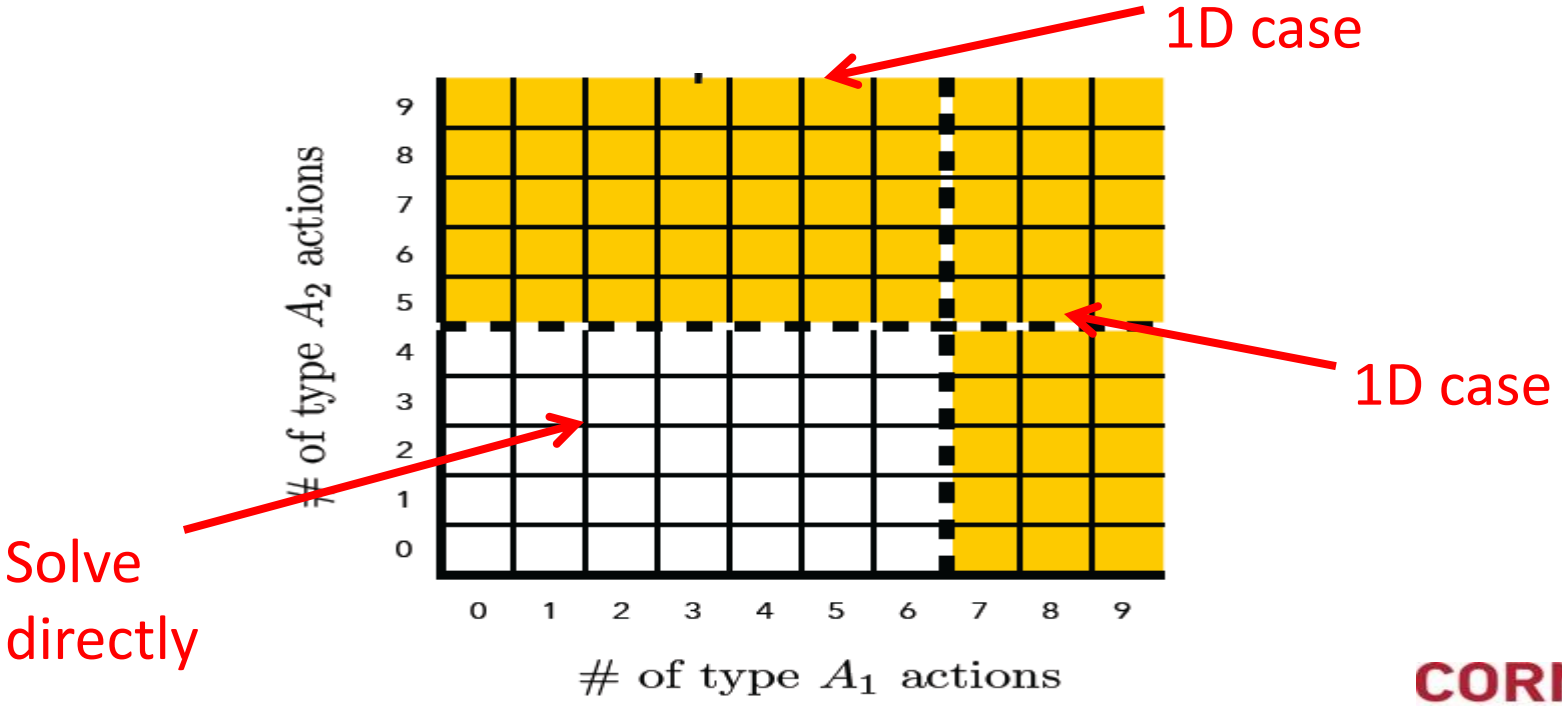
1D Example

- Level of targeted activity accelerates towards boundary



CORNELL NYCTECH

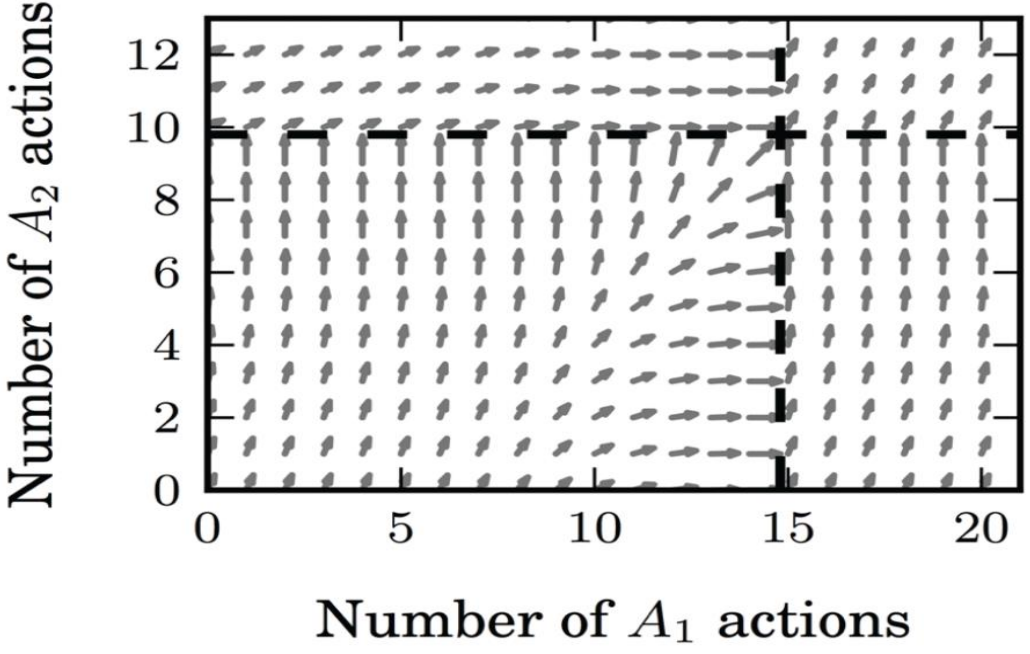
Optimization: Two Targeted Dimensions



CORNELL NYCTECH

2D Example

- Acceleration toward badge boundary



CORNELL NYCTECH

A Limitation of the Model

- Return to baseline, **p**, after achieving badges
- Does not allow for possible de-motivating effect of achieving a badge
- In practice an external incentive can lower a person's intrinsic incentive/preference
  - E.g., paying for blood donations can reduce the number of donors
- Possible extension, but not seen in our data

CORNELL
NYCTECH

StackOverflow Badges

- Extensive use of badges

- Consider two cumulative badges

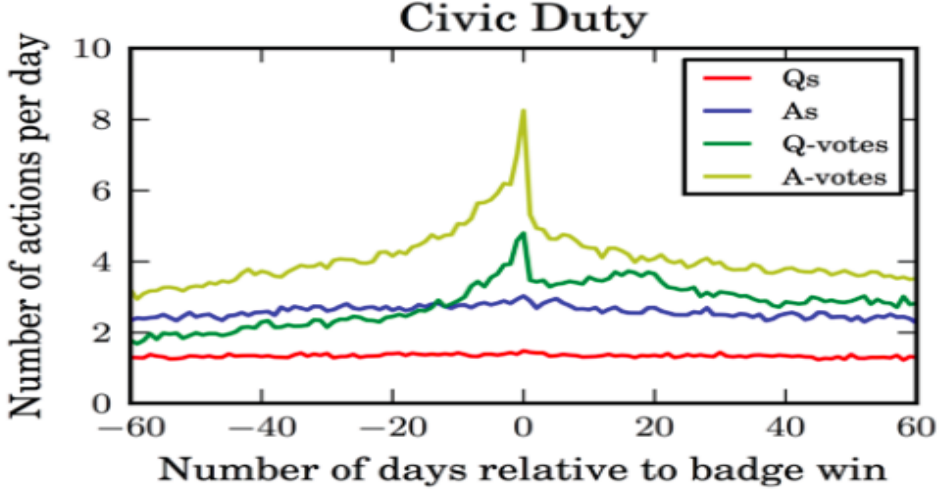
User votes on 600 questions

User votes 300 times

CORNELL NYCTECH

Civic Duty, 300 votes (Silver)

- Qualitatively consistent with model
  - Acceleration towards boundary
  - Increased targeted activity level and overall



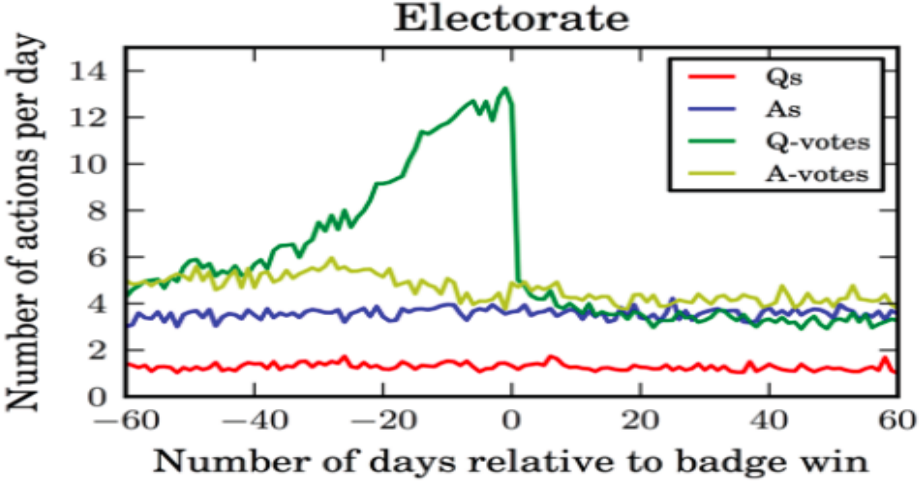
Note: aligned by day earned

For people active +/- 60 days from earning

CORNELL NYCTECH

Electorate, 600 Q Votes (Gold)

- Again qualitatively consistent
  - Single targeted dimension shows not only increased overall level but also tradeoff



CORNELL NYCTECH

Badge Placement Problem

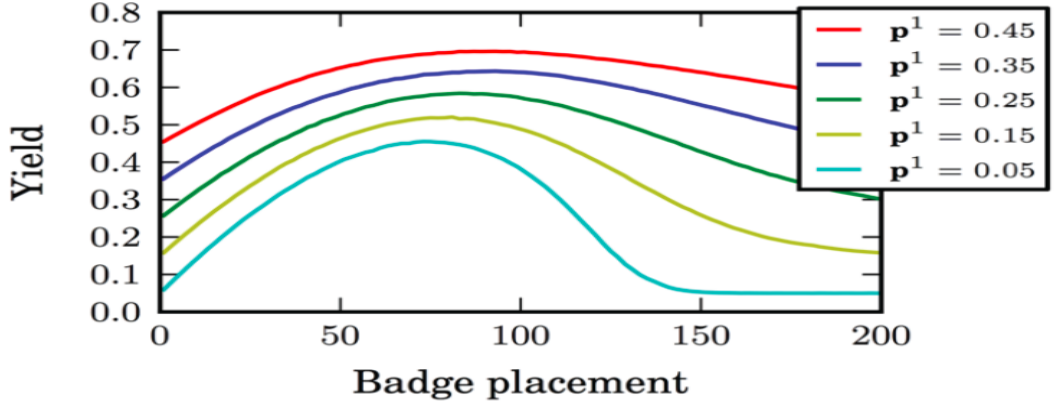
- How should designers “place” cumulative badges to achieve desired effects?
  - E.g., frequent flyer mile status levels, votes on StackOvervlow
- Define yield to be fraction of actions over lifetime on targeted dimension
  - Placement to maximize yield

User votes on ??? questions

CORNELL NYCTECH

Placing a Single Badge

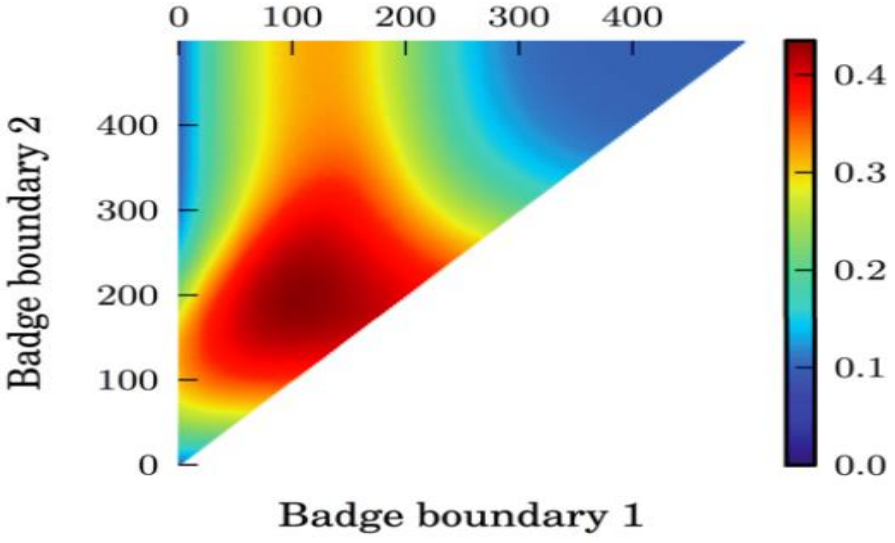
- Best yielding placements are those which are quite challenging for users to achieve
  - For Θ=.99 expect only 5 targeted actions for p^1=.05 yet optimal badge at 75



CORNELL NYCTECH

Two Badges on Single Dimension

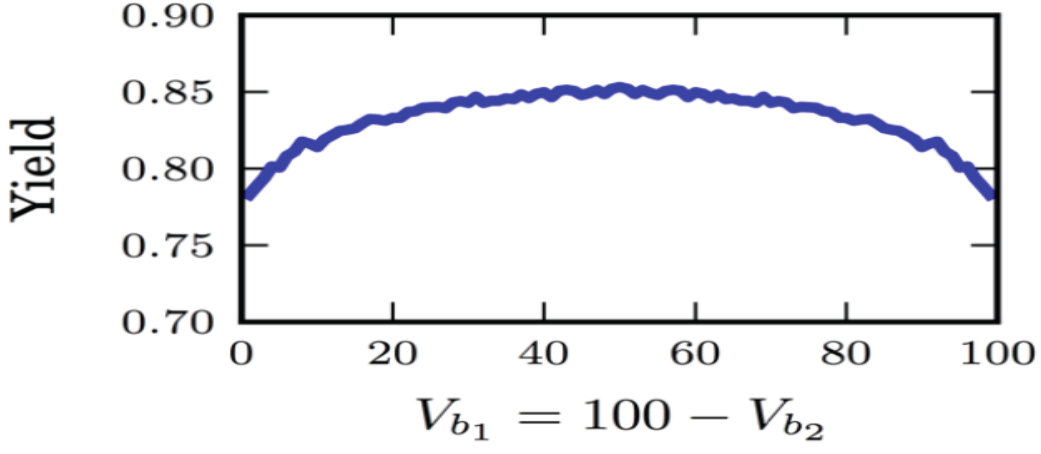
- Highest yield when badges placed relatively equally apart (illustration for Θ=.99)



CORNELL NYCTECH

Relative Badge Values

- For two badges with fixed total value, best to split value equally



CORNELL NYCTECH

Conclusions

- Introduced model of user behavior
  - Predicts users increase overall engagement and steer distribution actions to achieve badges
- Observe qualitative predictions in StackOverflow data
- Introduced badge placement problem
- Many questions
  - Where value in badges comes from
  - Competition and scarcity in badges
  - Analogies with offline domains
  - Badge system design

CORNELL NYCTECH