Steering User Behavior
With Badges

Dan Huttenlocher
Cornell Computer Science Department
and Cornell NYC Tech

Joint work with Ashton Anderson, Jon Kleinberg, Jure Leskovec
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’
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
Connected components in a graph with 100 million nodes

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.

1 Answer

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
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.
The Model

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

  Ask Q
  Answer Q
  Vote on Q
  .
  .
  .
The Action Space

- Action types $A_1$, $A_2$, ... form space denoting number of actions of each type
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:
  \[ g(p, p') = \| p - p' \|_2^2 \]
- User survives to next step with probability \( \Theta \) (generally \( \Theta=0.99 \)).
Badges

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

• User’s utility composed of 3 parts
  – Value from badges won
  – Cost of deviating from \( p \)

\[
f(a) = \sum_{b \text{ won}} V_b + \theta [p_a^1 \cdot f(a_1 + 1, a_2) + p_a^2 \cdot f(a_1, a_2 + 1)] - g(p, p_a)
\]

Expected utility of next state
Optimization: One Targeted Dimension

- Use dynamic programming to solve
- No reward for deviating from \( p \) past boundary
- Before boundary select \( p_a \) to maximize expected utility
- Collapses along A2 dimension
  - 1D problem, solve from boundary back to origin
1D Example

- Level of targeted activity accelerates towards boundary

Example: badge at 25 type $A_1$ actions
Optimization: Two Targeted Dimensions

Solve directly

1D case

# of type $A_2$ actions

# of type $A_1$ actions
2D Example

- Acceleration toward badge boundary
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
StackOverflow Badges

• Extensive use of badges

Peter Mortensen

5,214

6 27 62

• Consider two cumulative badges

• Electorate
  User votes on 600 questions

• Civic Duty
  User votes 300 times
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
Electorate, 600 Q Votes (Gold)

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

• How should designers “place” cumulative badges to achieve desired effects?
  – E.g., frequent flyer mile status levels, votes on StackOverflow

• Define yield to be fraction of actions over lifetime on targeted dimension
  – Placement to maximize yield

User votes on ??? questions
Placing a Single Badge

• Best yielding placements are those which are quite challenging for users to achieve
  – For $\Theta=0.99$ expect only 5 targeted actions for $p^1=0.05$ yet optimal badge at 75
Two Badges on Single Dimension

• Highest yield when badges placed relatively equally apart (illustration for Θ=.99)
Relative Badge Values

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

\[ V_{b_1} = 100 - V_{b_2} \]
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