Collective Intelligence

Tian Zheng

March 19, 2016

Collective Intelligence

[From wikipedia] shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals and appears in consensus decision making.

Web 2.0

[From wikipedia] World Wide Web sites that emphasize user-generated content.

Examples

- Rankings of products based on reviews
- Suggested friends based on your current friend list
- Ranking of scientific papers based on citations
- Suggested movies based on movies you have watched
- Sorting search results

Collective flitering

- Our online behaviors are being collected.
- Derive understanding on the latent `qualities` that drive our behaviors.
- Such understanding will guide recommendation engine to give recommendations.
- Including recommendations to ourselves.
- If we like everything the it recommended to us, does that mean it works?

Amazon movie reviews

- http://snap.stanford.edu/data/web-Movies.html
- Number of reviews 7,911,684
- Number of users 889,176
- Number of products 253,059
- Users with $>$ 50 reviews 16,341
- Median no. of words per review 101
- Timespan Aug 1997 - Oct 2012

Example

product/productId: B003AI2VGA

review/userId: A141HP4LYPWMSR

review/profileName: Brian E. Erland “Rainbow Sphinx”

review/helpfulness: 7/7

review/score: 3.0

review/time: 1182729600

review/summary: “There Is So Much Darkness Now ~ Come For The Miracle”

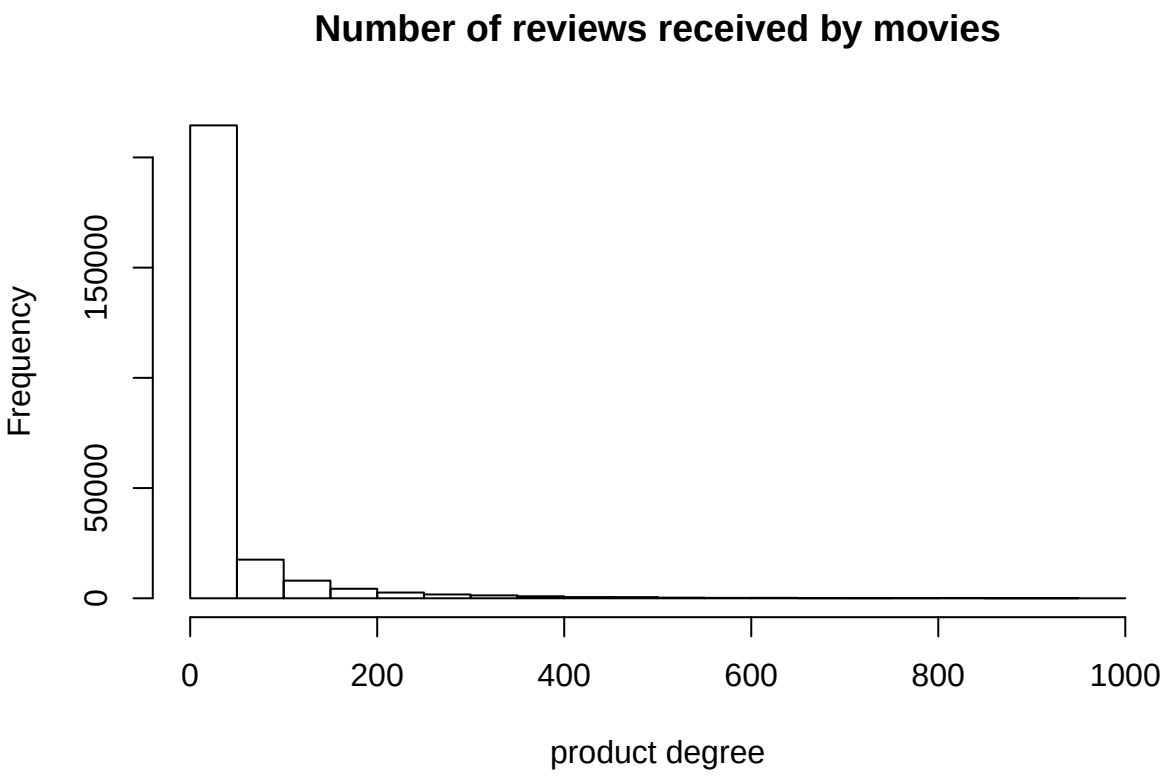
review/text: Synopsis: On the daily trek from . . .

Read in the data

- txt raw data was 8.7 GB.
- Very slow for R to process directly.
- Used Python to convert the file to CSV. (see the lib folder)
- For the purpose of this tutorial, I kept
- productId
- userId
- helpfulness
- score
- Output file: output/moiveiscsv.csv

Take a first look

Summary statistics



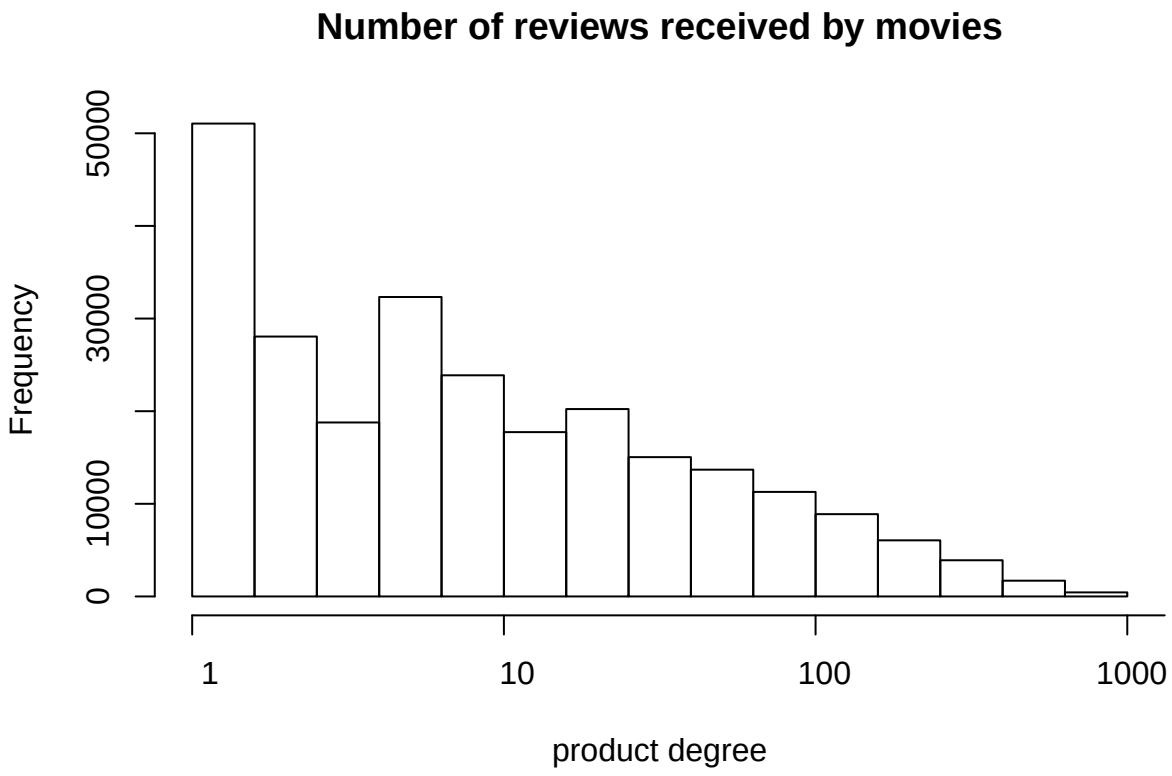

```
## B002QZ1RS6
##     200332
```

Summary statistics

```
## [1] 6
```

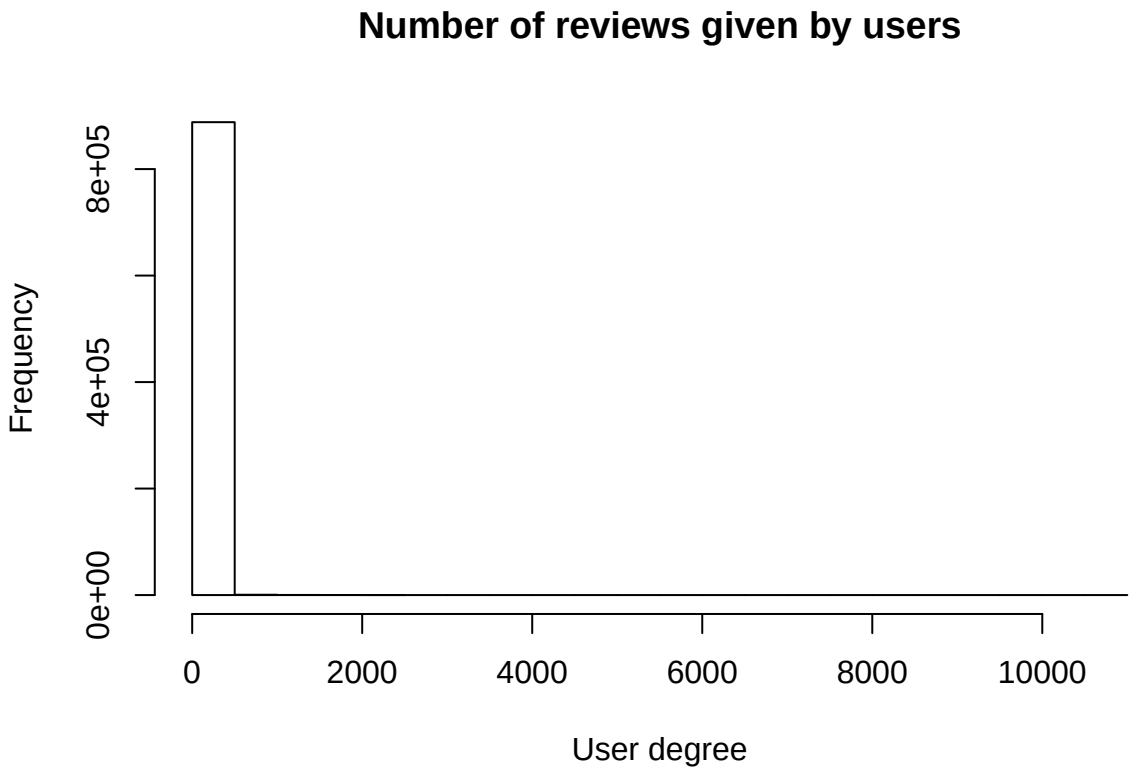
```
## [1] 31.26411
```

Summary statistics



Power-law distribution

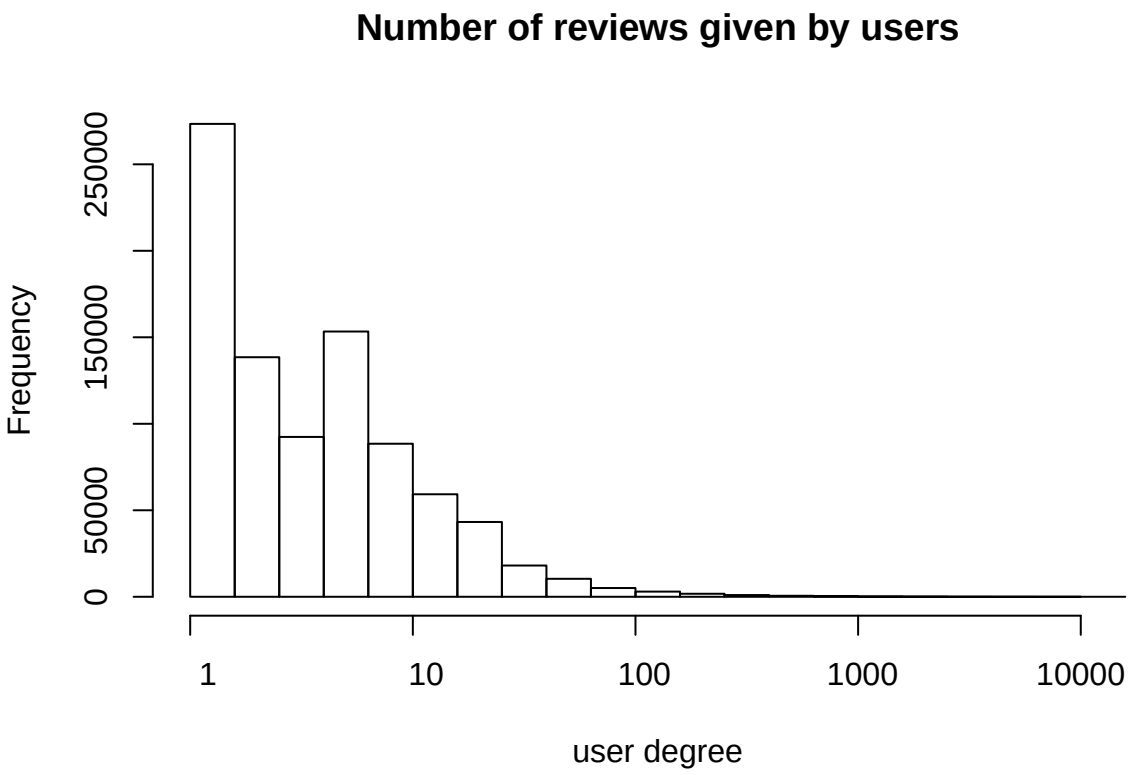
Summary statistics




```
## A2CRQHZTOP0TJI
##              3
```

```
## [1] 8.897774
```

Summary statistics



The reviews are more concentrated on a subset of users than movies.

Which are the best movies?

- simple average
- bayes approach
- normalized rating

Movie recommendations

Network analysis

Select movies and users with links.

```
##   product_productid  review_userid review_helpfulness review_score
## 1        B000063W1R A29DLKCN8QWO7B                1/1            5
## 2        B000063W1R A3QIEISBZP4QTV                1/1            5
## 3        B000063W1R A18758S1PUYIDT                1/1            4
```

Parsing the helpfuness votes

```
##         product_productid  review_userid review_helpfulness helpful.v
## 1311771      6304117302.0 A3CGYBFC5722PQ                1/1         1
## 2790489        B0028RXXFC  ADPVQ84PG7CT1                2/3         2
## 321865         B00006AG59   ATQWRIGPTGC2                0/0         0
```

```
##         total.v review_score  review_h
## 1311771       1            5 1.0000000
## 2790489       3            5 0.6666667
## 321865        0            4       NaN
```

Assortative mixing