Dennis Cornwell
CSE 881 Homework 1
9/8/13

1.a

- Top5Tweets.java
- Top5Followers.java

1.b. #Followee and #Tweets have the highest correlation with a value of 0.1936.

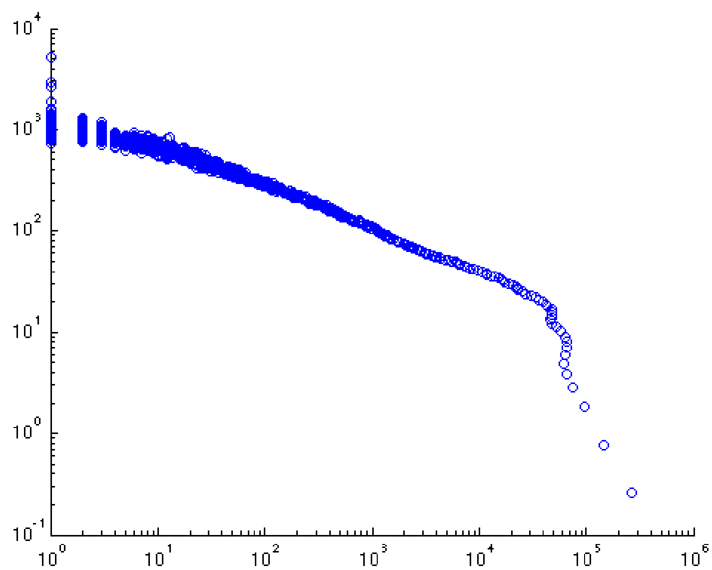
- (AggregateUserData.java for code to generate user.txt)

```
>> A = load('user.txt');
>> [r,p] = corrcoef( A(:,2:4) )
r =
    1.0000    0.0086    0.0217
    0.0086    1.0000    0.1936
    0.0217    0.1936    1.0000
```

- Script to generate scatter plots:

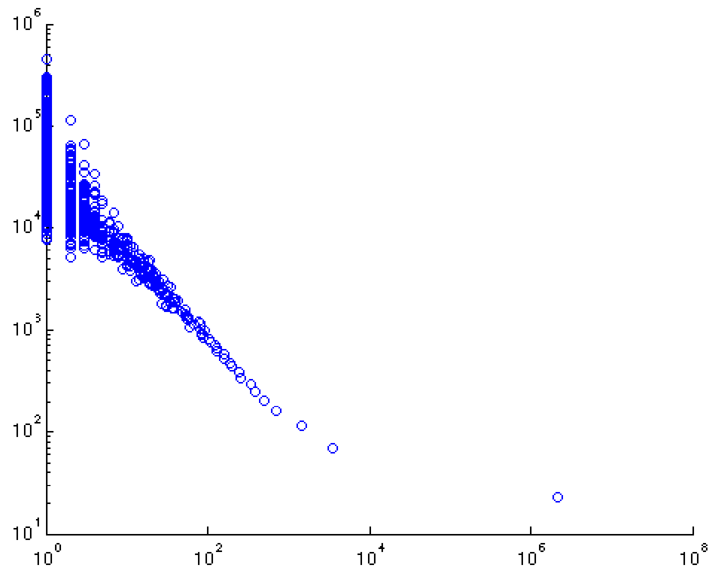
```
A = load('user.txt');
% Column 2 for followers, 3 for followees, 4 for tweets
[f,v] = hist( A(:,2), 10000 );
scatter(f,v);
set( gca, 'XScale', 'log' );
set( gca, 'YScale', 'log' );
```

Followees - Follows a power-law distribution

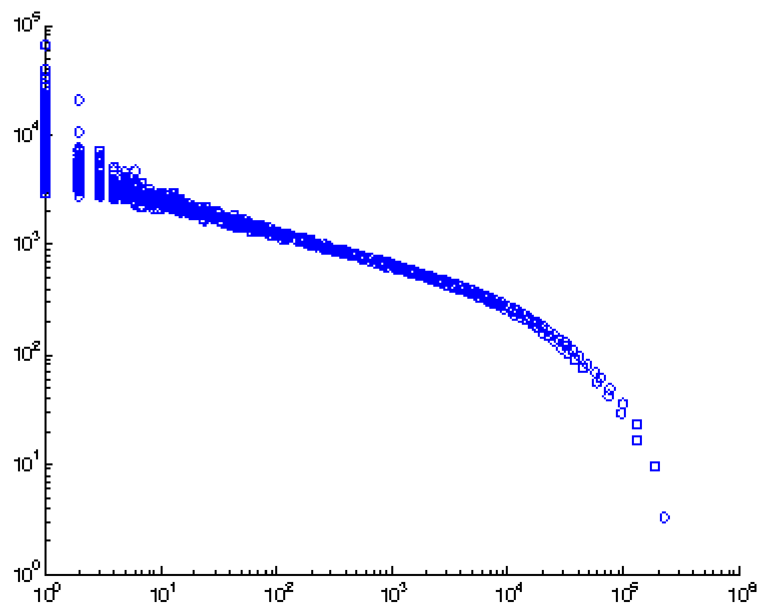


Dennis Cornwell


Followers - Does not follow a power law distribution




Tweets - Follows a power law distribution



Dennis Cornwell
CSE 881 Homework 1
9/8/13

2.

a. Orthogonal Projection:

   i. $Y - Y_x$ *perpendicular to* $\Omega$

   ii. $Y_x$ *projected into* $\Omega$

   i. *Let* $A = [X_1|X_2]$

   $(Y - Y_x)A = 0$

   $A^T Y = A^T Y_x$

   $A^T Y = A^T Y_x$

   ii. $Y_x$ *projected into the subspace* $\Omega$

   $Y_x = y_1 X_1 + y_2 X_2 = A\left|y_1; y_2\right|$, *Let* $\left|y_1; y_2\right| = Z$

   $Y_x = AZ$

   $A^T Y = A^T A Z$

   $(A^T A)^{-1} A^T Y = Z$

   $A(A^T A)^{-1} A^T Y = AZ$

   $A(A^T A)^{-1} A^T Y = Y_x$

b. $X_1 \times X_2 - \omega = 0 \Rightarrow \omega = X_1 \times X_2$

   and

   $||\omega||_2 = 1$

c.

   i. Using $A(A^T A)^{-1} A^T Y = Y_x$:

```
Yx =

    0.4381
    0.4190
    0.3238
```

   ii. $\omega_0 = X_1 \times X_2 = [0.02,\ 0.01,\ -0.04]$

   $||\omega|| = 1 :\ \omega = \frac{\omega_0}{||\omega_0||} = [0.4364,\ 0.218,\ -0.8729]$

3. Covariance

   a. No. Nothing in the question indicates that the covariance for the age vs. number of years in the community is normalized in such a way that it can be accurately compared with the covariance for height vs. number of years in the community, so a value of 5 compared with 0.5 does not necessarily mean that height is a better predictor of community membership duration.

b. using formula Cov(X,Y) E[XY] - E[X]E[Y]
   Let $a = (x - \bar{x})$ (conforming to the redefinition of x)
   Let $\bar{a} = \frac{x-\bar{x}}{n},\ E[x] = \bar{x}\,(for\ sufficiently\ large\ n)\ \Rightarrow \bar{a} = \frac{0}{n} = 0$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum(a-\bar{a})(y-\bar{y}) - \bar{a}\bar{y}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum(x-\bar{x}-0)(y-\bar{y}) - 0\bar{y}$

$\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y} < \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})\ assuming\ \bar{x}\ and\ \bar{y}\ may\ never\ take\ on\ negative\ values$

c. Let $a = cx$ (conforming to the redefinition of x where c is some constant)
   Let $\bar{a} = c\bar{x}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum(a-\bar{a})(y-\bar{y}) - \bar{a}\bar{y}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum(cx-c\bar{x})(y-\bar{y}) - c\bar{x}\bar{y}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum c(x-\bar{x})(y-\bar{y}) - c\bar{x}\bar{y}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{c}{n}\sum(x-\bar{x})(y-\bar{y}) - c\bar{x}\bar{y}$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y} < c(\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y})$

d. Let $a = \frac{x-\bar{x}}{\sigma_x}\ and\ b = \frac{y-\bar{y}}{\sigma_y}\ \Rightarrow\ 0 \le a, b \le 1$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum(a-\bar{a})(b-\bar{b}) - \bar{a}\bar{b}$

   Let $e = a - \bar{a}\ and\ f = b - \bar{b}\ with\ -1 \le e, f \le 1$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ ? =\ \frac{1}{n}\sum ef - \bar{a}\bar{b}$

   Since $-1 \le \frac{1}{n}\sum^{n} ef \le 1\ and\ 0 \le \bar{a}\bar{b} \le 1,$

   $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) - \bar{x}\bar{y}\ >\ \frac{1}{n}\sum(a-\bar{a})(b-\bar{b}) - \bar{a}\bar{b}$

4. Attribute Classification:
   a. Number of years since 1 BC - Discrete, Quantitative, Interval
   b. GPA received by a student - Discrete, Quantitative, Ordinal
   c. Mood of blogger - Discrete, Qualitative, Nominal
   d. Sound intensity in dB - Continuous, Quantitative, Ratio