Fitting the WHOIS Internet data

R. M. D’Souza†, C. Borgs∗, J. T. Chayes∗, N. Berger‡, and R. D. Kleinberg+†

†Dept. of Mechanical and Aeronautical Eng., University of California, Davis
∗Microsoft Research, Redmond, WA
‡Department of Mathematics, University of California, Los Angeles
+Department of Computer Science, Cornell University, Ithaca NY

This short technical manuscript contains supporting information for Ref. [1]. We consider the RIPE WHOIS internet data as characterized by the Cooperative Association for Internet Data Analysis (CAIDA) [2], and show that the Tempered Preferential Attachment (TPA) model [1] provides an excellent fit to this data. First we define the complementary cumulative probability distribution (ccdf), and then derive the ccdf for a TPA graph. Next we discuss the ccdf for the WHOIS data. Finally we discuss the fit provided by the TPA model and by a power law with exponential decay (PLED).

I. DEFINING THE CCDF

The complementary cumulative probability distribution, ccdf(x):

\[
\text{ccdf}(x) = 1 - \sum_{j=1}^{x-1} p_j = \sum_{j=x}^{\infty} p_j.
\]

II. THE CCDF PREDICTED BY TPA WITH \( A_1 \neq A_2 \)

A. First recall the recursion relations

The recursion relations defining the degree distribution for TPA graphs were derived explicitly in Refs. [3] and [4]. Here we derive the corresponding ccdf. These are Eqn’s (16) and (17) in [3]:

\[
p_i = \left( \prod_{k=2}^{i} \frac{k-1}{k+w} \right) p_1 = \left( \prod_{k=1}^{i-1} \frac{k}{k+w+1} \right) p_1, \quad \text{for} \quad i \leq A_2;
\]

and

\[
p_i = \left( \frac{A_2}{A_2 + w} \right)^{i-A_2} p_{A_2} = q^{i-A_2} p_{A_2}, \quad \text{for} \quad i \geq A_2.
\]

Note

\[
p_{A_2} = \left( \prod_{k=1}^{A_2-1} \frac{k}{k+w+1} \right) p_1,
\]
and, for convenience, we defined:

\[ q \equiv \left( \frac{A_2}{A_2 + w} \right). \quad (5) \]

We will first calculate the CCDF for \( i \geq A_2 \) as we will use that result to determine the CCDF for \( i < A_2 \).

**B. Calculating the CCDF, for \( x \geq A_2 \)**

Recall the definition of the CCDF from Eqn. (1):

\[
\text{ccdf}(x) = \sum_{j=x}^{\infty} p_j = p A_2 \sum_{j=x}^{\infty} q^{j-A_2} = p A_2 \sum_{j=0}^{\infty} q^{j+x-A_2} = p A_2 q^{x-A_2} \sum_{j=0}^{\infty} q^j. \quad (6)
\]

Since \( q < 1 \), the sum in Eqn. (6) is a geometric series; \( \sum_{j=0}^{\infty} q^j = 1/(1-q) \). Thus we can write:

\[
\text{ccdf}(x) = \left( \frac{p A_2}{1-q} \right) q^{x-A_2}, \text{ for } x \geq A_2. \quad (7)
\]

**C. Calculating the CCDF, for \( x < A_2 \)**

This is slightly more complicated, as we have different functional forms for \( x < A_2 \) and \( x > A_2 \).

\[
\text{ccdf}(x) = \sum_{j=x}^{\infty} p_j = \sum_{j=x}^{A_2-1} p_j + \sum_{j=A_2}^{\infty} p_j = \sum_{j=x}^{A_2-1} p_j + \text{ccdf}(A_2) = \sum_{j=x}^{A_2-1} p_j + \left( \frac{p A_2}{1-q} \right). \quad (8)
\]
Plugging in the relation for \( p_i \) from Eqn. (3), we obtain:

\[
\text{ccdf}(x) = p_{A_2} \left( \frac{1}{1 - q} + \frac{A_2 - 1}{A_2 - 1} \sum_{j=x}^{A_2 - 1} \prod_{k=j}^{A_2 - 1} \frac{k + w + 1}{k} \right), \quad \text{for } x < A_2.
\] (9)

D. Standard Normalization

First we can check that Eqns. (7) and (9) give the same value for \( \text{ccdf}(A_2) \). They do:

\[
\text{ccdf}(A_2) = \frac{p_{A_2}}{1 - q}. \] (10)

And we can determine the value of \( p_{A_2} \) by the normalization condition that

\[
\text{ccdf}(1) = 1 = p_{A_2} \left( \frac{1}{1 - q} + \sum_{j=1}^{A_2 - 1} \prod_{k=j}^{A_2 - 1} \frac{k + w + 1}{k} \right). \] (11)

In other words,

\[
p_{A_2} = \left( \frac{1}{1 - q} + \sum_{j=1}^{A_2 - 1} \prod_{k=j}^{A_2 - 1} \frac{k + w + 1}{k} \right)^{-1}. \] (12)

E. Normalizing without degree \( d = 1 \) nodes

We may want to neglect nodes with degree \( d < 2 \) for various reasons. In that case, the normalization would be:

\[
\text{ccdf}(2) = 1 = p_{A_2} \left( \frac{1}{1 - q} + \sum_{j=2}^{A_2 - 1} \prod_{k=j}^{A_2 - 1} \frac{k + w + 1}{k} \right). \] (13)

Thus

\[
p_{A_2} = \left( \frac{1}{1 - q} + \sum_{j=2}^{A_2 - 1} \prod_{k=j}^{A_2 - 1} \frac{k + w + 1}{k} \right)^{-1}. \] (14)

with Eqns. (7) and (9) unchanged (except Eqn. (9) now holds for \( 2 \leq x < A_2 \), rather than for \( 1 \leq x < A_2 \)).
III. THE WHOIS CCDF, FOR $d > 1$

A. Whois data, renormalize to remove $d < 2$

By definition:

$$\sum_{j=1}^{\infty} p_j = 1.$$  

Thus:

$$\sum_{j=2}^{\infty} p_j = 1 - p_1.$$  

We want to renormalize ($p'_j = \eta p_j$) such that:

$$\sum_{j=2}^{\infty} p'_j = \eta \sum_{j=2}^{\infty} p_j = 1,$$

Thus $\eta = 1/(1 - p_1)$. For the Whois data, $p_1 = 0.0573$. and $\eta = 1.0608$.

The complementary cumulative distribution function (ccdf) for the renormalized probabilities:

$$\text{ccdf}'(x) = \sum_{j=x}^{\infty} p'_j = \eta \sum_{j=x}^{\infty} p_j = \eta \text{ccdf}(x).$$

FIG. 1: Original CCDF of Whois data, and the renormalized CCDF$'(x) = \eta \text{CCDF}(x)$. 
IV. FITTING TPA TO WHOIS WITH $d \geq 2$

Whois $d \geq 2$ distribution discussed above. TPA with $d \geq 2$ is the same as with $d \geq 1$ except the value of $p_{A_2}$ is defined as in Eqn. (14), in terms of $d = 2$ instead of $d = 1$.

![Graph](image)

FIG. 2: Whois CCDF for $d \geq 2$. Data points are from the Whois tables. The solid line is the fit to TPA for $d \geq 2$ with $A_1 = 187$ and $A_2 = 90$ (and thus $\gamma = 1.83$). With this fit, $R = 0.986$, thus $R^2 = 0.972$.

V. FITTING PLED TO WHOIS WITH $d \geq 2$

Assuming a PLED: $p(x) = Ax^{-b}\exp(-x/c)$. The normalization constant, $A$, is determined by the relation:

$$\sum_{x=2}^{\infty} p(x) = 1 = A \sum_{x=2}^{\infty} x^{-b}\exp(-x/c).$$

Then the ccdf:

$$\text{ccdf}(x) = A \sum_{j=x}^{\infty} x^{-b}\exp(-x/c).$$
FIG. 3: Whois CCDF for $d \geq 2$. Data points are from the Whois tables. The solid line is the fit $ccdf(x) = A \sum_{j=x}^{\infty} x^{-b} \exp(-x/c)$, where $b = 1.63$ and $c = 350$. With this fit, $R = 0.985$, thus $R^2 = 0.970$.


