Evolutionary Forces Imprinted on the Network Structures of the Internet

Poong Oh and Peter Monge
June 17-18, 2014
A New Agenda for Internet Archives
Cambridge, MA
Primary Goals

• To develop a new theoretical model for network evolution by integrating the
  – *Moran process* in evolutionary biology (Moran, 1958, 1962)
  – *Vertex copying model* in network science (Kleinberg et al., 1999; Kumar et al., 2000)

• To detect *evolutionary forces* that have played out on the Web by applying the new model
  – *Scale-free networks* as null models (neutral drifts with no bounds)
  – *Maximal carrying capacity* of a network (ecological constrains)
  – *Selection* for replication (natural selection)

• To discuss *challenges* in using the Internet Archives data
  – Incompleteness of data
  – Accuracy of time-related information
Network Evolution?

• Commonly used in network literature since the publication of Erdős and Rényi’s (1960) “On the evolution of random graphs”
  – Network evolution as “gradual structural changes over time”

• Darwinian Evolution
  – “Strictly speaking, neither genes, nor cells, nor organisms, nor ideas evolve. Only populations can evolve” (Nowak, 2006, p. 26).
  – Populations of self-replicating entities with heritable traits.
  – Mutation occurs when entities fail to make perfect copies of themselves.
  – Variants are subjects to natural selection
  – Darwinian evolution occurs when any dynamical system has arisen which continually modifies itself by the three principles of replication, mutation, and selection
Moran Process for Evolutionary Dynamics
(Moran, 1958; Cressman, 2003; Nowak, 2006)
Moran Process for Evolutionary Dynamics
(Moran, 1958; Cressman, 2003; Nowak, 2006)

Time $t$

Select an individual for elimination at random

5:5
Moran Process for Evolutionary Dynamics
(Moran, 1958; Cressman, 2003; Nowak, 2006)

Select another for replication on a uniform random basis (i.e., neutral drift) or with a probability proportional to its fitness (i.e., natural selection)
Moran Process for Evolutionary Dynamics
(Moran, 1958; Cressman, 2003; Nowak, 2006)

Select another for replication on an uniform random basis (i.e., neutral drift) or with a probability proportional to its fitness (i.e., natural selection)
Moran Process for Evolutionary Dynamics
(Moran, 1958; Cressman, 2003; Nowak, 2006)

Time $t$

Select another for replication on an uniform random basis (i.e., neutral drift) or with a probability proportional to its fitness (i.e., natural selection)

5:5

Time $t+1$

Making an identical copy of itself

Replication

6:4

Failure to make a perfect copy

Mutation

5:4:1
Replicator Equation
(Maynard Smith, 1982; Cressman, 2003; Nowak, 2006)

- \( x_i \) is the relative frequency of type \( i \) in a population (e.g., the populations of the blue, red and green circles)
- At a given point in time, the growth rate of \( x_i \) can be written as

\[
\frac{dx_i}{dt} = \sum_{j}^{n} x_j f_j(x) \mu_{ji} - x_i \bar{f}
\]

where
- \( x = (x_1, x_2, x_3, ..., x_n) \): the vector of the relative frequencies of types in the population
- \( f_j(x) \): the \textit{fitness function} of type \( j \) (which is dependent on the population)
- \( \mu_{ji} \): the \textit{mutation rate} from type \( j \) to type \( i \)
- \( \bar{f} = \sum_{j}^{n} x_j f_j(x) \): the \textit{average fitness} of the entire population

- \textbf{Neutral drift}: \( f_i(x) = \bar{f} \) for all \( i \); no natural selection pressure

- By solving the replicator equation, \( f_i(x) \) and \( \mu_{ij} \) can be obtained, which enables us to capture the essence of evolutionary dynamics
  - See also the Price equation (Price, 1970) and the Lotka-Volterra equation (Lotka, 1925)
Extension of Moran Process

• The original Moran process assumes a fixed population size:
  – Guarantees equal numbers of the eliminated and the replicated
  – Implies “saturation,” i.e., consumption of all environmental resources
  – Focuses on equilibria rather than “equilibration” (Maynard Smith, 1982)

• Extended models involve changes in population size:
  – Growing population: \( n_e < n_r \)
  – Shrinking population: \( n_e > n_r \)
    where \( n_e \) and \( n_r \) are the numbers of the eliminated and the replicated, respectively.

• Possible growth functions of population:
  – Growth rate \( r \) = birth rate \( (n_r/N) \) – death rate \( (n_e/N) \)
  – Linear growth: \( N(t) = rt \)
  – Exponential growth: \( N(t) = N_0e^{rt} \)
  – Exponential growth to a limit: \( N(t) = K(1 - e^{-rt}) \)
  – Logistic growth: \( N(t) = \frac{KN_0e^{rt}}{K+N_0(e^{rt}-1)} \)
  – \( N_0 \) is the initial population; \( K \) is maximal carrying capacity
Vertex Copying Model
(Kleinberg et al., 1999; Kumar et al., 2000; Krapivsky & Redner, 2005)
Vertex Copying Model
(Kleinberg et al., 1999; Kumar et al., 2000; Krapivsky & Redner, 2005)

Time $t$

A new node arrives
Percent Copying Model
(Kleinberg et al., 1999; Kumar et al., 2000; Krapivsky & Redner, 2005)

Select at random an existing node whose linkage pattern is to be copied.
Vertex Copying Model
(Kleinberg et al., 1999; Kumar et al., 2000; Krapivsky & Redner, 2005)

Select at random an existing node whose linkage pattern is to be copied
Vertex Copying Model
(Kleinberg et al., 1999; Kumar et al., 2000; Krapivsky & Redner, 2005)

Time $t$

Time $t+1$

Replication
Making an identical linkage pattern

Mutation
Failure to make a perfect copy

poongoh@usc.edu
monge@usc.edu
Replicator Equation for Network Moran Process

- $x_i$ is the relative frequency of linkage pattern $i$ in a network
- At a given point in time, the growth rate of $x_i$ can be written as

$$\frac{dx_i}{dt} = \sum_{j}^{n} x_j f_j(A) \mu_{ji} - x_i \bar{f}$$

where
  - $A$: the adjacent matrix of the network
  - $f_j(A)$: the fitness function of linkage pattern $j$ (which is dependent on the network structure)
  - $\mu_{ji}$: the mutation rate from linkage pattern $j$ to $i$
  - $\bar{f} = \sum_{j}^{n} x_j f_j(A)$: the average fitness of the network \(\leftarrow\) growth function, $N(t)$

- Neutral drift: $f_i(A) = \bar{f}$ for all $i$; no natural selection pressure

- By solving the replicator equation, $f_i(A)$ and $\mu_{ij}$ can be obtained, which enables us to capture the essence of network evolution
Detecting Evolutionary Forces on Networks

• The Network Moran Process allows us to estimate the rates at which sub-populations of different linkage patterns grow over time (i.e., Darwinian fitness)

• Well defined “sub-population(s)” are required (mutually exclusive)
  – By nodes’ structural properties (e.g., degree, centrality, distance to a given node, etc.)
  – By nodes’ attributes (e.g., age, gender, prestige, role/function, etc.),
  – By sub-graphs’ structural properties (e.g., triad census), or
  – By any other criteria of interest and their combinations

• “Growth by replication” can be interpreted as “social learning through imitation”
  – Because individuals have only local information about the entire network, they cannot make “rational” decisions for tie formation (Weller, et al., 2010). Instead, they imitate the linkage patterns of “successful” others’ (Skyrms, 2010; see also Maynard Smith, 1982).
  – It is unnecessary to define “success” in advance, because “successful” nodes can be identified by their replication rates (i.e., ecological rationality, Boyd & Richerson, 2005)
Detecting Evolutionary Forces on Networks (Cont.)

• The growth rate of a type is jointly determined by

  – *Intrinsic growth rate* (e.g., the tendencies of reciprocity and transitivity in social networks)

  – Ecological constraints:
    • *Maximal carrying capacity:* the maximal numbers of nodes and edges that a network can sustain, Monge, et al., 2008).
    • As a network grows to its maximal carrying capacity, intrinsic growth rates are suppressed

  – Influences of others:
    • *Competitive (symbiotic) relations* with other types hamper (promote) one’s growth

  – *Mutation* rates:
    • Random errors in replication processes
    • Important because mutation allows a network to escape sub-optimum states, e.g., Innovative discoveries by “accidents”
Application to the network of .gov websites

Data collection

- A total of 200 websites of U.S. government departments and agencies
- From Sept. 1995 to present

<table>
<thead>
<tr>
<th>ID</th>
<th>Website</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U.S. AbilityOne Commission</td>
<td><a href="http://www.abilityone.gov">http://www.abilityone.gov</a></td>
</tr>
<tr>
<td>2</td>
<td>American Battle Monuments...</td>
<td><a href="http://www.abmc.gov">http://www.abmc.gov</a></td>
</tr>
<tr>
<td>3</td>
<td>Access Board</td>
<td><a href="http://www.access-board.gov">http://www.access-board.gov</a></td>
</tr>
<tr>
<td>4</td>
<td>Advisory Council on Historic...</td>
<td><a href="http://www.achp.gov">http://www.achp.gov</a></td>
</tr>
<tr>
<td>5</td>
<td>Administrative Conference...</td>
<td><a href="http://acus.gov">http://acus.gov</a></td>
</tr>
<tr>
<td>6</td>
<td>African Development Foundation</td>
<td><a href="http://www.adf.gov">http://www.adf.gov</a></td>
</tr>
<tr>
<td>200</td>
<td>The White House</td>
<td><a href="http://www.whitehouse.gov">http://www.whitehouse.gov</a></td>
</tr>
</tbody>
</table>
Application to the network of .gov websites

Network Construction

- Node: websites
- Link: $i \rightarrow j$, if any webpages of website $i$ contains hyperlinks to any webpages of website $j$.

<table>
<thead>
<tr>
<th>ID</th>
<th>Website</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U.S. AbilityOne Commission</td>
<td><a href="http://www.abilityone.gov">http://www.abilityone.gov</a></td>
</tr>
<tr>
<td>2</td>
<td>American Battle Monuments...</td>
<td><a href="http://www.abmc.gov">http://www.abmc.gov</a></td>
</tr>
<tr>
<td>3</td>
<td>Access Board</td>
<td><a href="http://www.access-board.gov">http://www.access-board.gov</a></td>
</tr>
<tr>
<td>4</td>
<td>Advisory Council on Historic...</td>
<td><a href="http://www.achp.gov">http://www.achp.gov</a></td>
</tr>
<tr>
<td>5</td>
<td>Administrative Conference...</td>
<td><a href="http://acus.gov">http://acus.gov</a></td>
</tr>
<tr>
<td>6</td>
<td>African Development Foundation</td>
<td><a href="http://www.adf.gov">http://www.adf.gov</a></td>
</tr>
<tr>
<td>200</td>
<td>The White House</td>
<td><a href="http://www.whitehouse.gov">http://www.whitehouse.gov</a></td>
</tr>
</tbody>
</table>

For example,

- Administrative Conference...
  - http://acus.gov/
- African Development...
  - http://www.adf.gov/

For example, poongoh@usc.edu

monge@usc.edu
Application to the network of .gov websites

Network Visualization

Federal Aviation Administration
http://www.faa.gov
1995/09/05

In-degree = 141
Out-degree = 6

The White House
www.whitehouse.gov
In-degree = 141
Out-degree = 6

2001/01/01

2005/01/01

2009/01/01

2014/06/01
Application to the network of .gov websites

Growth of the Network
Application to the network of .gov websites

Growth of the Network

![Graph showing exponential growth to a limit](image)

**Exponential growth to a limit:**

\[ N = K(1 - e^{-rt}) \]

where \( K \) is the upper limit, and \( r \) is the intrinsic growth rate

\[ \hat{K} = 240.19 [240.14, 240.24] \]
\[ \hat{r} = 0.0236 [0.0232, 0.0240] \]
\[ \text{Adj. } R^2 = 0.985 \]
Application to the network of .gov websites

Growth of the Network

Exponential growth to a limit:

\[ N = K(1 - e^{-rt}) \]

where \( K \) is the upper limit, and \( r \) is the intrinsic growth rate.

\[ \hat{R} = 240.19 \quad [240.14, 240.24] \]
\[ \hat{r} = .0236 \quad [.0232, .0240] \]
\[ \text{Adj. } R^2 = .985 \]

Member Carrying Capacity:
The maximal number of nodes that a network can sustain

(Monge, et al., 2008)
Application to the network of .gov websites
Defining sub-populations by linkage pattern

• Triad: a sub-network that consists of three nodes
• Triad census (Davis & Leinhardt, 1972)
  – All triads in a network can be classified into 16 types:
Application to the network of .gov websites
Defining sub-populations by linkage pattern

• Triad: a sub-network that consists of three nodes
• Triad census (Davis & Leinhardt, 1972)
  – All triads in a network can be classified into 16 types:

**Open triads:**
Maximizing the efficiency of information flow by reducing redundant links (Burt, 1992)

**Closed (transitive) triads:**
A general tendency of social networks;
“A friend of my friend is likely to be my friend” (Wasserman & Faust, 1994)
Application to the network of .gov websites

Competitive growth of open and closed triads

Relative Frequency of Closed Triads ($p_c$)

Time ($t$)

poongoh@usc.edu
monge@usc.edu
Application to the network of .gov websites

Competitive growth of open and closed triads

Significantly larger than those of random graphs:
m = 12.25%; sd = 7.91
95% CI: [12.12, 12.38]
(Wasserman, 1977)

Relative Frequency of Closed Triads ($p_c$)

Will “closed triads” be dominant in this network?
Application to the network of .gov websites
Competitive growth of open and closed triads
**Application to the network of .gov websites**

Competitive growth of open and closed triads

Logistic growth:

\[ p_c(t) = \frac{K}{1 + e^{rt}} \]

where \( K \) is the upper limit, and \( r \) is the intrinsic growth rate

\[ \hat{R} = 0.250 \quad [0.246, 0.254] \]
\[ \hat{r} = 1.010 \quad [0.991, 1.029] \]
\[ \text{Adj. } R^2 = 0.964 \]
Application to the network of .gov websites
Competitive growth of open and closed triads

Non-linear growth dynamics of two types

The network is optimized by keeping balance between the two conflicting forces.

Survival of many: Neither of two types can take over the entire population. Instead, they coexist.

Logistic growth:

\[ p_c(t) = \frac{K}{1 + e^{rt}} \]

where \( K \) is the upper limit, and \( r \) is the intrinsic growth rate.

\[ \hat{K} = .250 \ [0.246, 0.254] \]

\[ \hat{r} = 1.010 \ [0.991, 1.029] \]

Adj. \( R^2 = .964 \)

Relative Frequency of Closed Triads (\( p_c \))

Time (\( t \))

poongoh@usc.edu
monge@usc.edu
Challenges in Using IA Data

Censoring problems

Webpage A

birth

Tie formation from B to A

Tie formation from A to B

Tie Resolution

Death

Time ($t$)

Webpage B

birth

Death

Time ($t$)

poongoh@usc.edu
monge@usc.edu
Challenges in Using IA Data

Censoring problems

Webpage A

- Birth
- Snapshots
- Tie formation from B to A
- Tie resolution
- Death

Webpage B

- Birth
- Snapshots
- Interval-censoring
- Death

Time ($t$)
Conclusions

• The network Moran process captures non-linear growth dynamics of different linkage patterns as a network grows over time.

• By doing so, it reveals underlying mechanisms of network evolution.
  – Ecological constraints
  – Intrinsic growth rates of different types
  – Competitive and symbiotic relations among them
  – Mutation rates

• Also, it allows us to make inferences about future states of networks.
  – The finding suggests that the maximal number of .gov websites is about 240.
  – What would be the maximal number of websites on the whole Internet?