COMS W4995-1: Intro to Social Networks Fall 2012

Handouts #2 – Reinforcement process

Instructor: *Augustin Chaintreau* Teaching Assistant: Zhikun Ma, Christopher Riederer

The particular case which we will study has been originally motivated by the appearance of species inside a genus (in classification of plants and animals), which were found to obey also power law distribution. We can see that new specie or genus can appear by mutation and the key to this phenomenon is that a genus containing more species is more likely to see mutation occuring among them.

Dynamics of a simple balls and bins reinforcement process The following dynamics have been introduced by G. Yule in 1925 to model evolution of species and their repartition within different classes, or "genus". We model a species as a ball which is contained in a bin that model its associated class or genus. Once a specied appear it is never removed and it does not change genus (*i.e.*, once a ball is created, it remains forever in the same bin). The main dynamics is the apparition of new species.

We assume that the process is initialized at time $t = 0$ with an arbitrary number of genera containing each an arbitrary number of species. The time follows discrete time slots denoted $t = 1, 2, \ldots$, and we introduce the following notation:

$$\forall t \geq 0, \forall i \geq 0\,,\ X_i(t) \text{ is the number of genera containing exactly } i \text{ species.}$$

Hence the distribution of species among genus at time t is entirely described by the sequence $(X_i(t))_{i\geq 1}$. In particular,

We assume that during a time slot, a mutation occurs in exactly one species, that is chosen uniformly at random among all species. The consequence of this mutation is the following:

- With a probability $p > 0$ (chosen independently of the past), this mutation is so important that it creates a new genus by itself. As a consequence, a new bin is created that contains exactly one ball. All other bins remains unchanged.
- Otherwise (and hence with probability $(1 - p)$), this mutation generates a new species which is contained in the same genus as the original one that was chosen. As a consequence, a new ball is created in the bin that contained the original ball that was chosen for the mutation.

Intuitively, in the second case, the process entails *reinforcement* for the following reason: in a genus that contains twice more species, the chance of having a mutation is twice more likely and hence this genus is also twice more likely to "grow" by having another species created and associated with it. In other words, the larger a genus becomes, the faster it will grow.

The following result proves formally that this reinforcement implies that as time grows large, the number of species contains by a genus approaches a power law.

Theorem 1 (Analysis of Yule Process)**.** *For the dynamics described above,*

(*i*) *There exist $C_1, C_2, \ldots$ such that for any $i \geq 1$ we have, $X_i(t)/t \to C_i$ almost surely as t grows large.*

(*i*) *We have $C_1 = \frac{p}{2-p}$ and $C_i = C_{i-1}(1 - a/i + O(i^{-2}))$ where $a = \frac{2-p}{1-p}$*

(i) *this implies* $\ln(\frac{C_i}{C_1}) \sim -\alpha \ln(i)$ *which implies that* C_i *is roughly proportional* i^{-a} *and hence is well approximated by a power law.*

Proof. The proof follows from three steps

1. Evolution equation for the expected value of $X_i(t)$.
2. Analysis of the limit $t \to \infty$ of this evolution
3. A probablistic concentration result and its consequence

Note that the third argument is provided for your information, but it is out of scope of the actual topic of this course. A supplementary question in the assignment deals with it in case you feel courageous enough to manipulate this argument.

Step1: Evolution equation for the expected value of $X_i(t)$: Let $X_1(t)$ be the number of genus with exactly 1 species, we wish to prove $\frac{X_1(t)}{t} \to C_1$.

Let us first translate the dynamics of the system in a given step into the evolution of the variable X_1 between time slots t and $t+1$.

Let us denote by $N(t)$ the number of species in the system. Note that since during each time slot a mutation occurs and create exactly one species, we have $N(t) = N(0) + t$.

We have:

$$X_1(t+1) = \begin{cases} X_1(t)+1 & \text{with probability } p \\ X_1(t)-1 & \text{with probability } (1-p)\frac{X_1(t)}{N(t)} \\ X_1(t) & \text{otherwise} \end{cases}$$

The first case follows from the fact that with probability a new genus is created with exactly one species. The second case represents the case when the mutation occurs in one of the $X_1(t)$ species associated with genera containing a single species, and it is not creating a new genera. Indeed in this case, one of these genera will contain two species in the next time slot. Finally the last case denotes any other event.

Finding the Expected Value of $X_1(t)$

As a consequence of the previous dynamics we can very precisely characterize the evolution of the *expectation* of $X_1(t)$ with time t:

That is to multiply the probability of each event by the each case which are defined above

$$E[X_1(t+1)] = p*E[X_1(t)+1] + \frac{E[X_1(t)]}{N(t)}*(1-p)*E[X_1(t)-1] + \{1-p-\frac{E[X_1(t)]}{N(t)}*(1-p)*E[X_1(t)]\}$$
$$E[X_1(t+1)] = E[X_1(t)] + p - \frac{E[X_1(t)]}{N(t)}*(1-p)$$

How about $X_i(t)$ and its expected value?

Let $X_i(t)$ be a number of genus with exactly i species. We wish to show $\frac{X_i(t)}{t} \to C_i$

$$X_i(t+1) = \begin{cases} X_i(t)+1 & \text{with probability } (i-1)\frac{X_{i-1}(t)}{N(t)}(1-p) \\ X_i(t)-1 & \text{with probability } \frac{iX_i(t)}{N(t)}(1-p) \\ X_i(t) & \text{otherwise} \end{cases}$$

The same principle applies for finding the expected value for $X_i(t)$, that is to multiply the probability of each event.

$$E[X_i(t+1)] = E[X_i(t)] + \frac{(i-1)E[X_{i-1}(t)]}{N(t)}(1-p) - \frac{iE[X_i t)]}{N(t)}(1-p)$$

Step2: Limit of expected value of $X_i(t)/t$ Let $\Delta_i(t) = E[X_i(t)] - t \cdot C_i$, we want $\frac{\Delta_i(t)}{t} = 0$ as t goes to ∞.

Let us first prove it for $i = 1$. We want to prove that $\Delta_1(t)$ is small According to the above evolution of $\mathbb{E}[X_1(t)]$ we have:

$$\begin{aligned}\Delta_1(t+1) &= \Delta_1(t) - C_1 + E[X_1(t+1)] - E[X_1 t)] \\ &= \Delta_1(t) - C_1 + p - \tfrac{\Delta_i(t)+t\cdot C_i}{N(t)}(1-p)\end{aligned}$$

Putting all factors of $\Delta_1(t)$ together, we obtain

$$= \Delta_1(t)[1 - \tfrac{1-p}{N(t)}] \underbrace{-C_1 + p - t \cdot \frac{C_1(1-p)}{N(t)}}_{\frac{(-C_1+p)N(t)-C_1(1-p)t}{N(t)}}$$

, where $N(t)$ is nothing but $N(0) + t$

the underbraced expression may be rewritten as: $\frac{(-C_1+p)N(0)+t[-C_1+p-C_1(1-p)]}{N(t)}$

As t grows large, $N(t)$ grows large as well. In order to show that this term becomes small, we wish to have the coefficient multiplying t to be zero. That is we assume:

$-C_1 + p - C_1(1-p) = 0 \Longrightarrow C_1 = \frac{p}{2-p}$.

Note that we can assume that since, until now, the value of the constant C_1 was not fixed and can be chosen arbitrarily for all these results to hold.

Now we have that the underbraced expression reduces to a term becoming small as t grows:

$\Delta_1(t+1) = \Delta_1(t)[1 - \frac{1-p}{N(t)}] + \frac{(-C_1+p)N(0)}{N(t)}$

Since the term multiplying $\Delta_1(t)$ is less than 1 in absolute value and the right term is less that $\frac{A}{t}$ for a constant $A > 0$ chosen sufficiently large, we can apply Lemma 2 below

$|\Delta_1(t+1)| \leq \Delta_1(0) + A \sum_{s=1}^{t} \frac{1}{s}$

Hence, using Lemma 3, we deduce that there exists $A' > 0$ such that

$\Delta_1(t) \leq A' \ln(t)$, which proves in particular that $\frac{\Delta_1(t)}{t}$ goes to zero as t gets large.

We wish to prove the following hypothesis for any $i \geq 1$: $\forall \varepsilon > 0\ \exists A$ such that $(|\Delta_i(t)| \leq A t^{\varepsilon})$,

Indeed, we have just shown that this is true for $i = 1$ since we found a logarithmic upper bound on the size of $\Delta_1(t)$. By recurrence, it is sufficient to prove that if it holds for $i-1$ it holds for i as well.

Note that, following similar steps as used for $i = 1$ (rewriting evolution of expectation $\mathbb{E}[X_i(t+1)]$ using $\Delta_i(t)$ and $\Delta_{i-1}(t)$, we have:

$$\Delta_i(t+1) = \Delta_i(t)\left(1 - \frac{i(1-p)}{N(t)}\right) + \frac{(i-1)(1-p)}{N(t)}\Delta_{i-1}(t) + \underbrace{\left(-C_i + \frac{(i-1)(1-p)t\cdot C_{i-1}}{N(t)} - \frac{i(1-p)t\cdot C_i}{N(t)}\right)}_{=\frac{N(0)C_i + t\cdot\left(-C_i+(1-p)(i-1)C_{i-1}-(1-p)iC_i\right)}{N(t)}}.$$

Note that, again the value of C_i is not fixed so that we can choose C_i so that the coefficient of t in the underbraced term is zero. This implies:

$$(1-p)(i-1)C_{i-1} = C_i + (1-p)iC_i$$

or, in other words $C_i = C_{i-1}\left(1 - \frac{2-p}{1+(1-p)i}\right) = C_{i-1}\left(1 - \frac{2-p}{(1-p)i} + \underbrace{\frac{2-p}{(1-p)i(1+(1-p)i)}}_{\leq \frac{A}{i^2} \text{ for any constant } A \text{ larger than } \frac{2-p}{(1-p)^2}}\right)$.

Once this is shown, for any $\varepsilon > 0$, we can use the hypothesis for $i-1$ to deduce that $\Delta_{i-1} \leq At^{\varepsilon}$ and hence

$$\Delta_i(t+1) = \Delta_i(t)\gamma_t + S_t$$

where $|\gamma_t| < 1$ and $|S_t| \leq (i-1)(1-p)At^{\varepsilon-1} + A'/t \leq A''t^{\varepsilon-1}$. We can then conclude, using lemma 4 below that the hypothesis remains true for i.

Lemma 2. *If $\chi_{n+1} = \gamma_n\chi_n + S_n$, where $|\gamma_n| \leq 1$, then we have $|\chi_n| \leq |\chi_0| + \sum_{m=1}^{n} |S_m|$.*

Lemma 3. *For any $n \geq 1$ we have: $\sum_{j=1}^{n} \frac{1}{j} \leq 1 + \ln(n)$*

Lemma 4. *For any $\varepsilon > 0$ and $n \geq 1$ we have: $\sum_{j=1}^{n} j^{\varepsilon-1} \leq 1 - \frac{1}{\varepsilon} + \frac{1}{\varepsilon}j^{\varepsilon}$.*

Step 3: ingredient probabilistic concentration result. So far, we have been able to prove that there exists $C_1, C_2, \ldots$ such that, as t grows, the expectation of $X_1(t)$, $X_2(t)$, $\ldots$ grows approximately as $C_1 \cdot t$, $C_2 \cdot t$ etc. (*i.e.*, for all $i \geq 1$, we have $\lim_{t\to\infty} \frac{\mathbb{E}[X_i(t)]}{t} = C_i$). We now wish to establish a much stronger result, comparable to a law of large number, which states that the sequence of random variables $X_i(t)$ grows approximately as $C_i \cdot t$ (i.e $\frac{X_i(t)}{t}$ converges to C_i, with probability 1).

To prove this we use the following probabilistic concentration result which states that, as t goes large, the sequence $X_i(t)$ is not far from its average value. More precisely, we admit the following result:

$$\forall i \geq 1\,,\ \forall M{>}0,\ \text{ we have } P[|X_i(t) - E[X_i(t)]|{>}M] \leq 2 * \exp(-\frac{M^2}{8t})\,. \tag{1}$$

If we choose M $= \sqrt{t * \ln(t)} * 4$

$$\underbrace{\frac{-M^2}{8t}}_{} = \frac{-16t\ln(t)}{8t} = -2\ln(t)$$

$$\exp(\underbrace{\frac{-M^2}{8t}}) = \exp(-2lnt) = \exp(\ln(t^{-2})) = \tfrac{1}{t^2}$$

We then have $P[|X_1(t) - E[X_i(t)]| >$M$] = \frac{1}{t^2}$

Then we can say,

$\implies P[|X_i(t) - E[X_i(t)]| >M] \leq \frac{1}{t^2}$
$\implies \sum_{t=0}^{\infty} P[|X_i(t) - E[X_i(t)]| >M] \leq \sum \frac{1}{t^2} < \infty$

Let us now conclude. For any $i \geq 1$ using Lemma 5 below (known as Borel Cantelli lemma), we know that there exists T which is almost surely finite such that when $t \geq T, |X_i(t) - E[X_i(t)]| \leq M$, which also implies that $|\frac{X_i(t)}{t} - \frac{E[X_i(t)]}{t}| \leq \frac{M}{t} = 4\sqrt{\frac{\ln(t)}{t}}$

Since $\lim_{t\infty} \frac{E[X_i(t)}{t} = C_i$ (note that this is a deterministic convergence for a sequence of real numbers). This means that almost surely (*i.e.*, on the event $\{T < \infty\}$), the values taken by the sequence of random variable $\frac{X_i(t)}{t}$ form a sequence of real numbers which are converging to a convergence sequence.

This implies that these values form a sequence that converge to the same limit, and hence that almost surely (*i.e.*, on the event $\{T < \infty\}$), $\frac{X_i(t)}{t}$ converges to C_i.

Lemma 5. *Let $(A_n)_{n\geq 0}$ be a sequence of event satisfying $\sum P(A_n)<\infty$.*
There exists N which is finite almost surely such that for any $n \geq N$, A_n does not occur.

Finally, let us prove that C_i is well approximated by a power law with coefficient $a = \frac{2-p}{1-p}$:
We have seen that $C_i = C_{i-1}(1 - \frac{\alpha}{i} + O(\frac{1}{i^2}))$

Note that this implies that there exists $A > 0$ such that $C_{i-1}(1-\frac{\alpha}{i}-\frac{A}{i^2}) \leq C_i \leq C_{i-1}(1-\frac{\alpha}{i}+\frac{A}{i^2})$
Which may be rewritten $C_1 \prod_{j=1}^{i}(1 - \frac{\alpha}{j} - \frac{A}{j^2}) \leq C_i \leq C_1 \prod_{j=1}^{i}(1 - \frac{\alpha}{j} + \frac{A}{j^2})$

This implies $\ln(C_1) + \sum_{j=1}^{i} \ln(1 - \frac{\alpha}{j} - \frac{A}{j^2}) \leq \ln(C_i) \leq \ln(C_1) + \sum_{j=1}^{i} \ln(1 - \frac{\alpha}{j} + \frac{A}{j^2})$

Hence $\exists A' > 0$ such that $\ln(C_1) + \sum_{j=1}^{i} \frac{\alpha}{j} - \sum_{j=1}^{i} \frac{A'}{j^2}) \leq \ln(C_i) \leq \ln(C_1) - \sum_{j=1}^{i} \frac{\alpha}{j} + \sum_{j=1}^{i} \frac{A'}{j^2})$

Note that $\alpha \sum_{j=1}^{i} \frac{1}{j} \sim -\alpha \ln(i)$ as i grows. This implies the result as the two other series are convergent and hence bounded.

If we assume that this approximation is exact we obtain
$\implies \ln(C_i) = \ln(C_1) - \alpha * \ln(i)$
$\implies C_i = C_1 * i^{-\alpha}$
$\implies C_i \propto i^{-a}$

This equality is not true in general (since the approximation introduces a constant that may play a role after being exponentiated). But one can show that there exist two constants A, B such that $Ai^{-\alpha} \leq C_i \leq Bi^{-\alpha}$, so that C_i remains not too far from a power law with the corresponding exponent. □