Ανάλυση Κοινωνικών Δικτύων (Social Network Analysis)

3η Εργαστηριακή Άσκηση

Συμεών Παπαβασιλείου (papavass@mail.ntua.gr)
Βασίλειος Καρυώτης (vassilis@netmode.ntua.gr)
Κωνσταντίνα Σακκά (nsakka@cn.ntua.gr)
Μαργαρίτα Βιτοροπούλου (mvitoropoulou@netmode.ntua.gr)
Γιώργος Μήτσης (gmitsis@netmode.ntua.gr)
Κωνσταντίνος Τσιτσεκλής (ktsitseklis@netmode.ntua.gr)

OneMax Problem

Βρείτε τη **δυαδική** ακολουθία (x1, x2,…, xn) που μεγιστοποιεί το άθροισμα $x_1 + x_2 + \ldots + x_n$, n=20.

- **Ποιότητα της λύσης: πόσο κοντά είναι η λύση του γενετικού αλγορίθμου στην προφανή βέλτιστη λύση**.
- **Πληθυσμός** από *10* έως *200* χρωμοσώματα με βήμα *10*.
- **Πιθανότητα διασταύρωσης** από *0.3* έως *0.9* με βήμα *0.1.*
- **Πιθανότητα μετάλλαξης** από *0.01* ως *0.2* με βήμα *0.01*.
- Εξετάστε την ποιότητα της λύσης του γενετικού αλγορίθμου **για κάθε συνδυασμό** μεγέθους πληθυσμού, πιθανότητας μετάλλαξης και πιθανότητας διασταύρωσης που προκύπτει.

Application in Community Detection

Social Network: G=(V,E)

represented by an Adjacency Matrix: $A=[a_{ij}]$

Problem:

Find communities in the social network ⟹ Find a partitioning of *A* into sub-matrices.

Let $S = (I, J)$ be sub-matrix of A, where I is a subset of the rows $\{I_1, \ldots, I_N\}$ of A, and J is a subset of the columns $\{J_1, \ldots, J_N\}$ of A.

Let a_{iJ} denote the *mean value* of the ith row of the S, and a_{Ij} the mean of the jth column of S.

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}, \text{ and } a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$$

The *volume* v_S of a sub-matrix $S = (I, J)$ is the number of 1 entries a_{ij} such that $i \in I$ and $j \in J$, that is $v_S = \sum_{i \in I, j \in J} a_{ij}$.

Application in Community Detection

Given a sub-matrix $S = (I, J)$, the *power mean of S of order r*, denoted as $\mathbf{M}(S)$ is defined as

$$\mathbf{M}(S) = \frac{\sum_{i \in I} (a_{iJ})^r}{|I|}$$

Definition of the fitness function

The *score* of S is defined as $Q(S) = \mathbf{M}(S) \times v_S$.
The *community score* of a partitioning $\{S_1, \ldots S_k\}$ of A is defined as

$$\mathcal{CS} = \sum_{i}^{k} Q(S_i)$$

Εντοπισμός Κοινοτήτων σε Γράφους Κοινωνικών Δικτύων με Χρήση Γενετικών Αλγορίθμων

Αναπαράσταση υποψήφιων λύσεων (χρωμοσωμάτων)

Κάθε χρωμόσωμα $b_{i,t}$ του πληθυσμού τη χρονική στιγμή t είναι ένα διάνυσμα με μήκος ίσο με το πλήθος των κόμβων του δικτύου, n. Αν $b_{i,t}(j) = \mathrm{d}$ τότε οι κόμβοι j, d ανήκουν στην ίδια κοινότητα. Επιπλέον για να ισχύει $b_{i,t}(j) = \mathrm{d}$, οι κόμβοι j, d θα πρέπει να συνδέονται στην αρχική τοπολογία. Τότε, βρίσκοντας τις συνδεδεμένες συνιστώσες του χρωμοσώματος $b_{i,t}$, προκύπτουν οι κοινότητες του δικτύου.

Αρχικοποίηση

Η αρχική γενιά του πληθυσμού θα πρέπει να αποτελείται από χρωμοσώματα τα οποία θα είναι τυχαία επιλεγμένα και διορθωμένα ώστε ο αριθμός που υπάρχει σε κάθε θέση να είναι γείτονας του αντίστοιχου κόμβου. (Αν δεν ισχύει θα πρέπει να αντικατασταθεί από έναν γείτονα του αντίστοιχου κόμβου.) Θεωρείστε πληθυσμό 300 χρωμοσωμάτων.

Συνάρτηση Fitness

Η συνάρτηση fitness υπολογίζεται όπως στις διαφάνειες της αντίστοιχης διάλεξης του μαθήματος. Προσοχή για κάθε συνιστώσα (component) του γράφου θα πρέπει να βρίσκετε τον αντίστοιχο υπογράφο.

Εντοπισμός Κοινοτήτων σε Γράφους Κοινωνικών Δικτύων με Χρήση Γενετικών Αλγορίθμων

Επιλογή (Selection)

Η επιλογή χρωμοσωμάτων από μία γενιά του πληθυσμού για την κατασκευή της επόμενης γενιάς γίνεται με τη μέθοδο της ρουλέτας (διαφάνειες μαθήματος και αλγόριθμος). Τροποποιήστε τον παραπάνω αλγόριθμο ώστε να εφαρμόσετε και ελιτισμό, δηλαδή τα x πρώτα χρωμοσώματα της νέας γενιάς θα είναι εκείνα τα χρωμοσώματα της προηγούμενης γενιάς που έχουν επιτύχει τη μέγιστη τιμή της συνάρτησης fitness.

Διασταύρωση (Crossover)

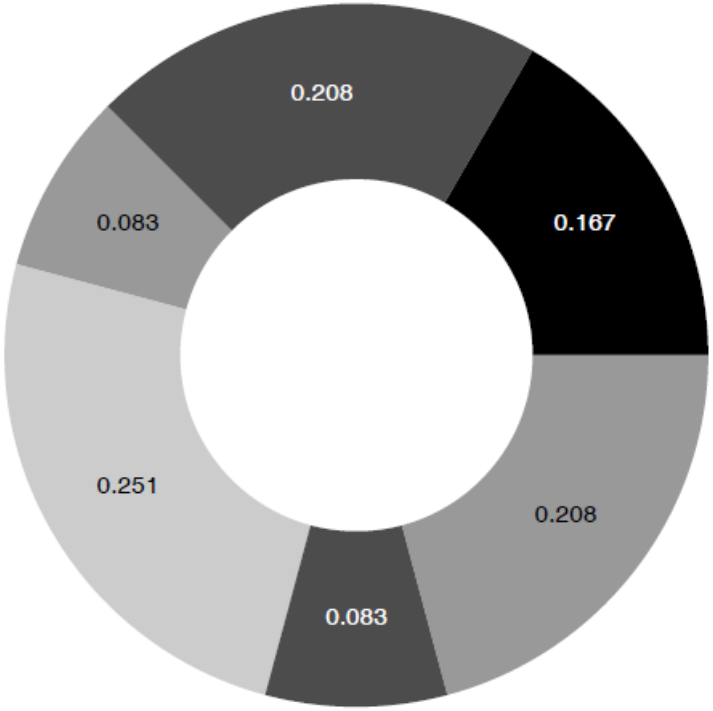
Η διασταύρωση πραγματοποιείται όπως περιγράφεται στον αλγόριθμο, μεταξύ διαδοχικών χρωμοσωμάτων της εκάστοτε γενιάς. Η πιθανότητα διασταύρωσης ανά ζεύγος είναι ίση με p_C.

Μετάλλαξη (Mutation

Η μετάλλαξη πραγματοποιείται σε κάθε θέση του χρωμοσώματος με πιθανότητα p_M, όπως στον αλγόριθμο. Θα πρέπει ωστόσο ο αλγόριθμος να τροποποιηθεί ώστε να δίνει ένα τυχαίο γείτονα του κόμβου που αντιστοιχεί σε κάθε θέση ως αποτέλεσμα της μετάλλαξης.

Selection – Algorithmic Description

- Generalized roulette game: in a roulette game, the slots are not equally wide, i.e. the different outcomes can occur with different probabilities.



Select an individual – chromosome as in the scheme of the previous slide.

Algorithm 1: Selection process

```
1 x := Random[0, 1];
2 i := 1;
3 while i < m & x > Σ_{j=1}^{i} f(b_{j,t}) / Σ_{j=1}^{m} f(b_{j,t}) do
4     i := i + 1;
5 select b_{i,t};
```

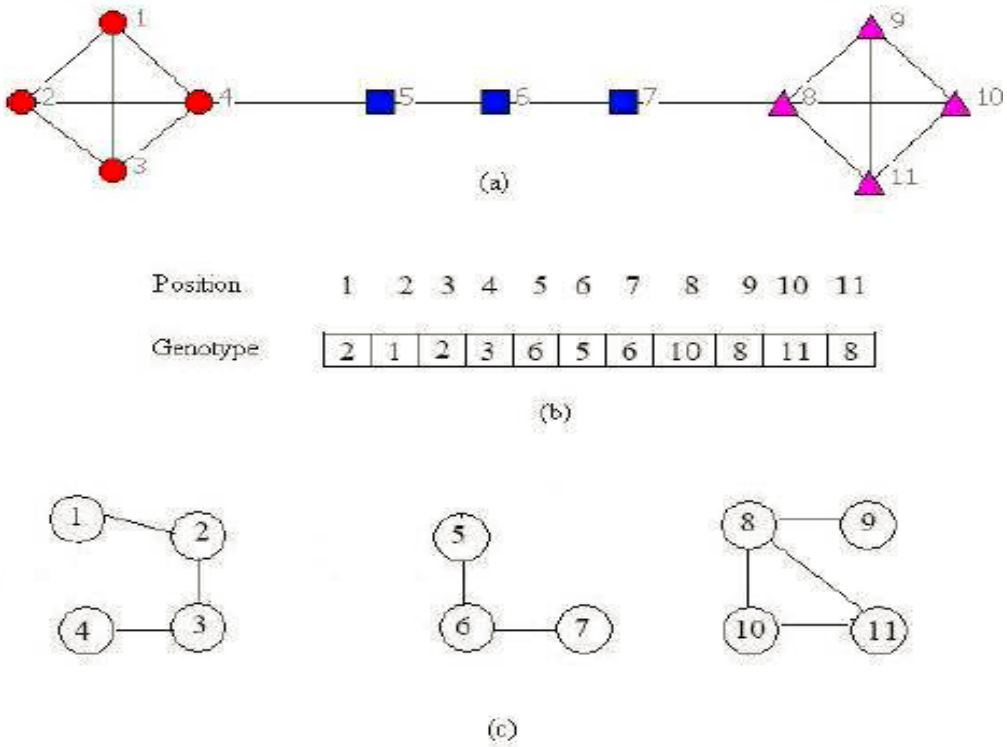
Application in Community Detection

Representation:

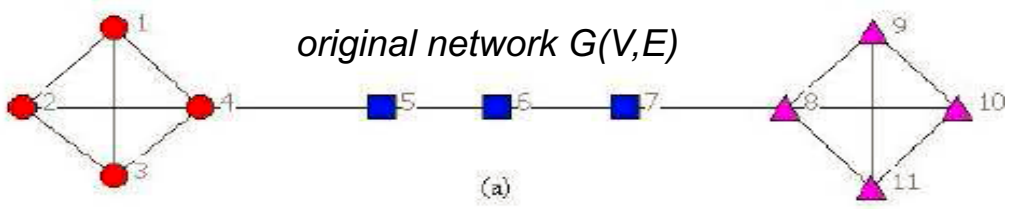
- Each individual is a vector of size *N* (number of nodes), where the value *j* at position *i* means a link between nodes *(i,j).*
- Such links should exist in the original network.
- Initialization via “safe” individuals i.e. check and correct if the links do not exist. If the value at position *i* is *k* and the link *(i,k)* does not exist, replace this link with a neighbor of *i*.
- Find communities represented by each individual via finding connected components.

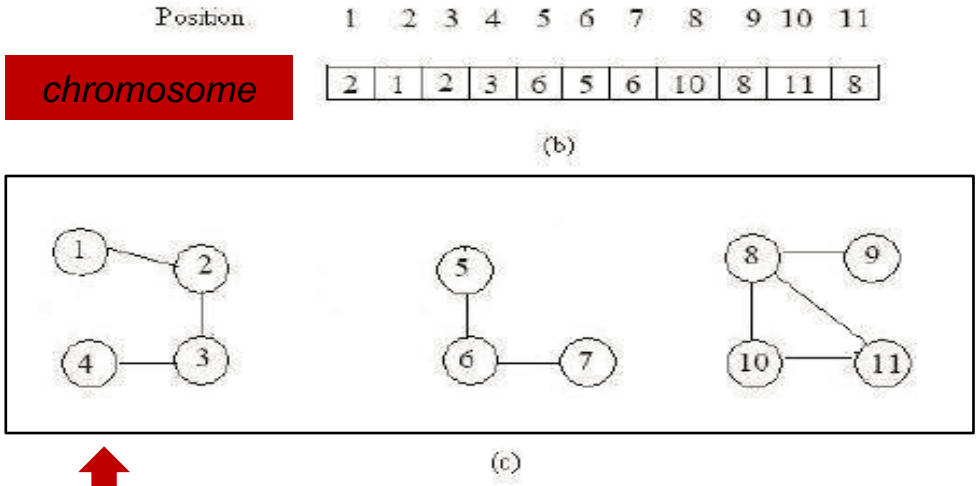
Mutation

- At position *i* only with the neighbors of *i*.



Application in Community Detection





We use the above partition, produced by the chromosome, to find out which nodes belong in each connected component (community).

The chromosome determines a partition of the original network into 3 components Ci, i=1,2,3:
C1 with the nodes {1,2,3,4},
C2 with the nodes {5,6,7} and
C3 with the nodes {8,9,10,11}
The adjacency matrix A of the original network G(V,E), V={1,2,…,11} is thus partitioned to the submatrices S1, S2, S3 with:

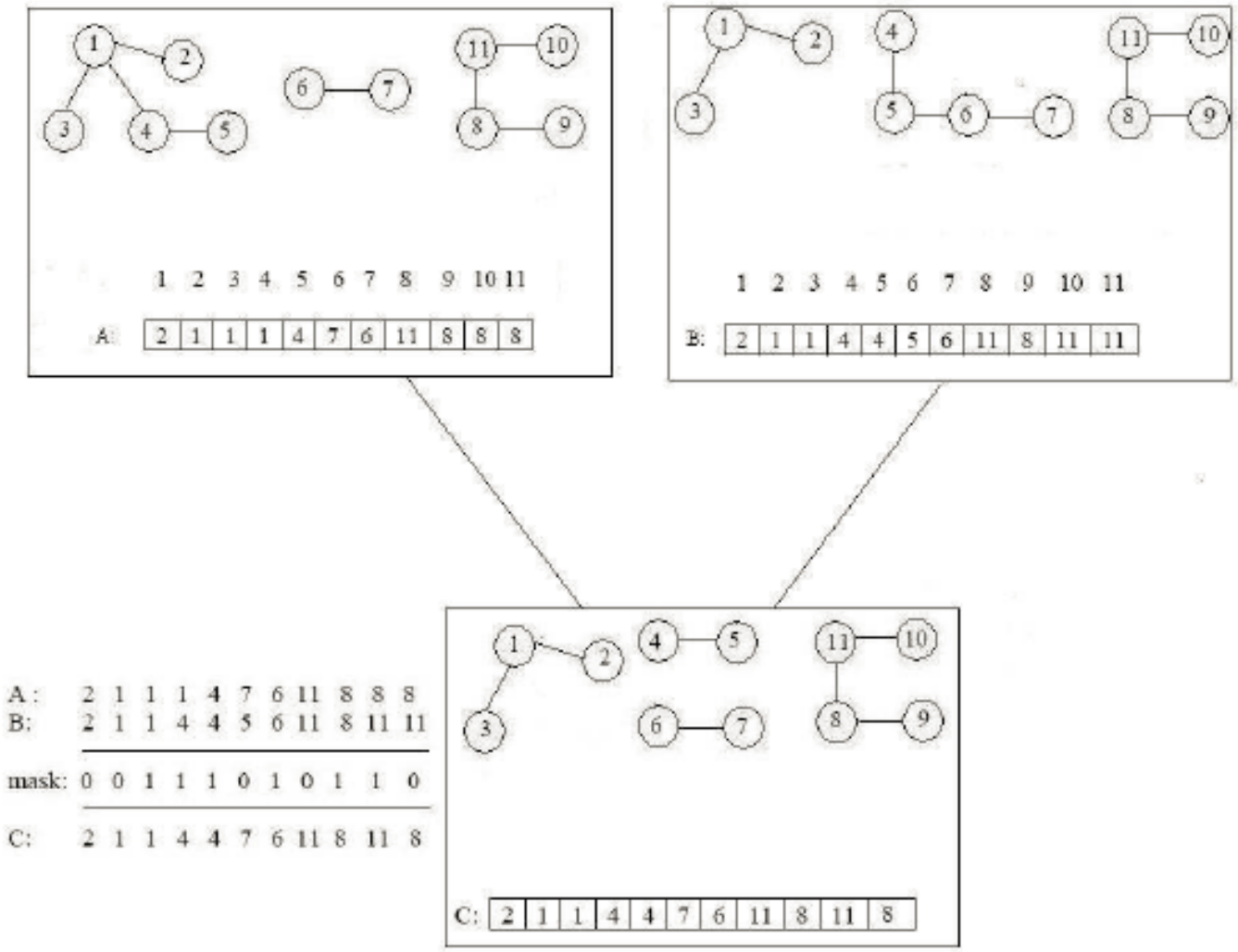
$$S1 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \qquad S2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$S3 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Attention: the edges between the nodes of the components are determined by the original network!

Application in Community Detection

Uniform Crossover: exchange randomly genes of safe parents.



Κριτήριο Τερματισμού (Stopping Criterion)

Ως κριτήριο τερματισμού ορίστε ένα μέγιστο πλήθος επαναλήψεων εφαρμογής του αλγορίθμου (γενεών πληθυσμού), ίσο με 30, ή ότι η βέλτιστη τιμή της συνάρτησης fitness δεν έχει μεταβληθεί για τις τελευταίες 5 επαναλήψεις.

Εφαρμόστε τον παραπάνω γενετικό αλγόριθμο στις πραγματικές τοπολογίες της δεύτερης εργαστηριακής άσκησης που δίνονται και στον Πίνακα 1. Οι παράμετροι που θα πρέπει να εξεταστούν δίνονται στον παρακάτω πίνακα. Προσοχή θα πρέπει να εξεταστούν όλοι οι συνδυασμοί.

Παράμετρος	**Τιμές**
Διασταύρωση p_C	0.7:0.9 με βήμα 0.1
Μετάλλαξη p_M	0.1, 0.2
Ελιτισμός x	1:3 με βήμα 1

Εντοπισμός Κοινοτήτων σε Γράφους Κοινωνικών Δικτύων με Χρήση Γενετικών Αλγορίθμων

- Συγκρίνετε το αποτέλεσμα του γενετικού αλγορίθμου **(το καλύτερο για τις διάφορες επιλογές παραμέτρων)** με εκείνο των αλγορίθμων εντοπισμού κοινοτήτων της εργαστηριακής άσκησης 2 **υπολογίζοντας το modularity.**

- **ΠΡΟΣΟΧΗ!** Θα πρέπει να έχετε πρωτίστως μετατρέψει το αποτέλεσμα του γενετικού αλγορίθμου σε κατάλληλη μορφή για να θεωρηθεί ως είσοδος στην modularity.

SIR Model

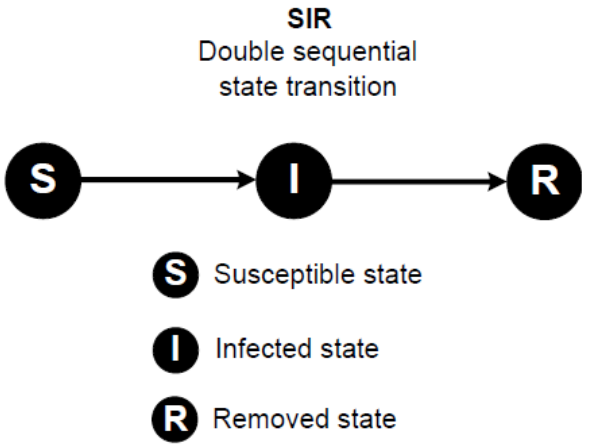
- **SIS: Susceptible – Infected - Removed**
  - Also Kermack-McKendrick model
- Models specific & single threats
- 2 state transitions for nodes
  - S to I to R
- System of differential equations needed
  - γ is the mean recovery (healing) rate

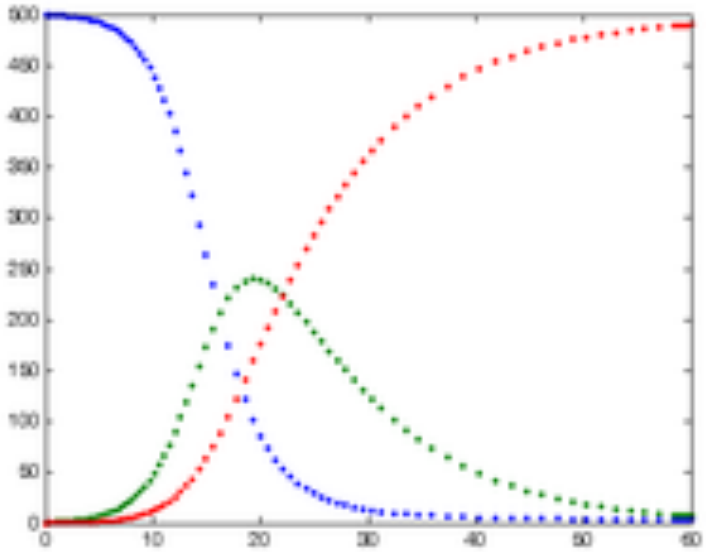
$$\frac{dS(t)}{dt} = -\frac{\beta S(t) I(t)}{N}$$

$$\frac{dI(t)}{dt} = \frac{\beta S(t) I(t)}{N} - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

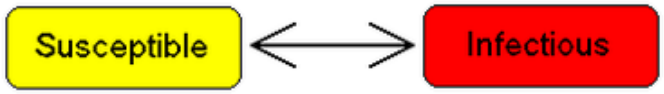
$$S(t) + I(t) + R(t) = N$$

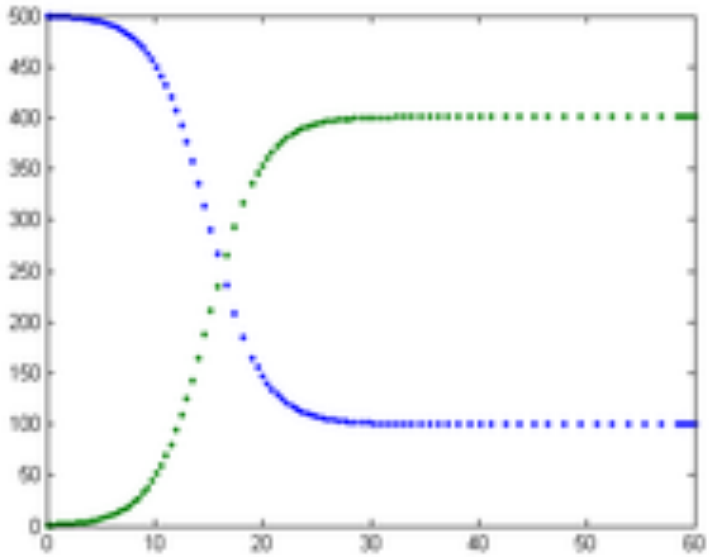




SIS Model

- **SIS: Susceptible – Infected - Susceptible**
- Examples are common cold & flu
- Macroscopic malware modeling
- No immunity after recovery





Based on the following set of equations

$$\frac{dI}{dt} = (\beta(t)N - \alpha)I - \beta(t)I^2$$

$$\beta(t) = 2 - 1.8\cos(5t)$$

$$\frac{dS}{dt} + \frac{dI}{dt} = 0 \rightarrow S(t) + I(t) = N$$

Modified (periodic) contact rate – e.g. flu/common cold