Data Mining: Data

Lecture Notes for Chapter 2---Data

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Transformation（变换）

- Similarity [0,1] scale
  - 1,2…10　　s’=(s-1)/9

相似度：$s' = (s - \min\_s)/(\max\_s - \min\_s)$

相异度：$d' = (d - \min\_d)/(\max\_d - \min\_d)$

- [0, ∞]　　$d' = d/(1+d)$
  - 0, 0.5, 2, 10, 100, 1000
  - 0, 0.33, 0.67, 0.90, 0.99, 0.999

Transformation（变换）

- Similarity to dissimilarity
  - 0,   1,   10,   100   (dissimilarity)
  - 1,   0.5,   0.09,   0.01 $s = 1/(1+d)$
  - 1.00, 0.37, 0.00, 0.00 $s = e^{-d}$
  - 1.00, 0.99, 0.00, 0.00 $s = 1 - \frac{(d - \min\_d)}{(\max\_d - \min\_d)}$
  -

Similarity/Dissimilarity for Simple Attributes

具有单个序数属性的对象（数据点）

- {poor, fair, OK, good, wonderful}

  - {poor=0, fair=1, OK=2, good=3, wonderful=4}

$$d=(x-y)/(n-1)$$

  -

Euclidean Distance

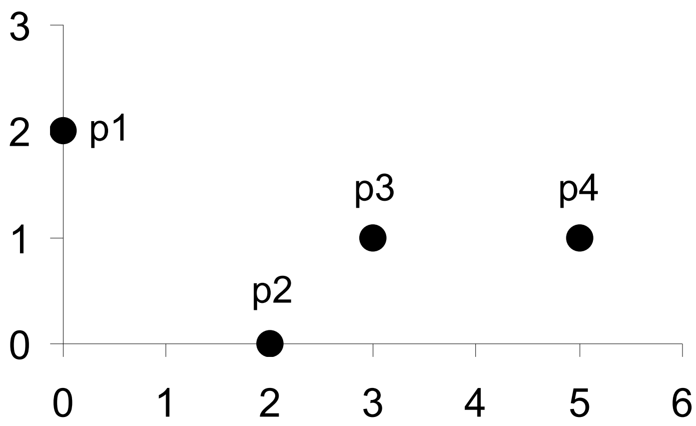
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k[th] attributes (components) of data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance 闵可夫斯基距离

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects p and q.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance
- $r = 2$. Euclidean distance
- $r \to \infty$. "supremum(上确界)" (L_{max} norm, L_∞ norm) distance
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

$$dist = \lim_{r \to \infty} \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}} = \max \left| p_k - q_k \right|$$

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_{∞}	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

  1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
  2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
  3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

  where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

  1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

  2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

  where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of “11” matches / number of not-both-zero attributes values

  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

$p =$ 1 0 0 0 0 0 0 0 0 0

$q =$ 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

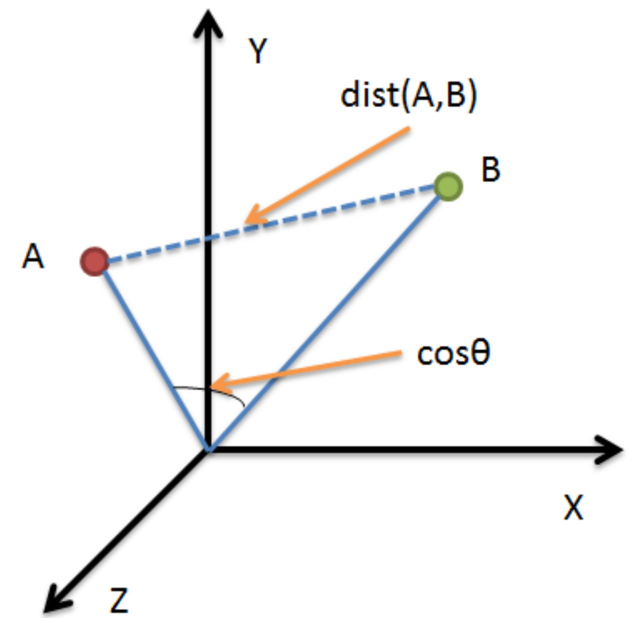
Cosine Similarity

- If d_1 and d_2 are two document vectors, then

  $$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

  where $\bullet$ indicates vector dot product

  and $|| d ||$ is the length of vector d.



- Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$$

$$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{\mathbf{0.5}} = (6)^{\mathbf{0.5}} = 2.245$$

$$\cos(d_1, d_2) = 0.3150$$

Correlation

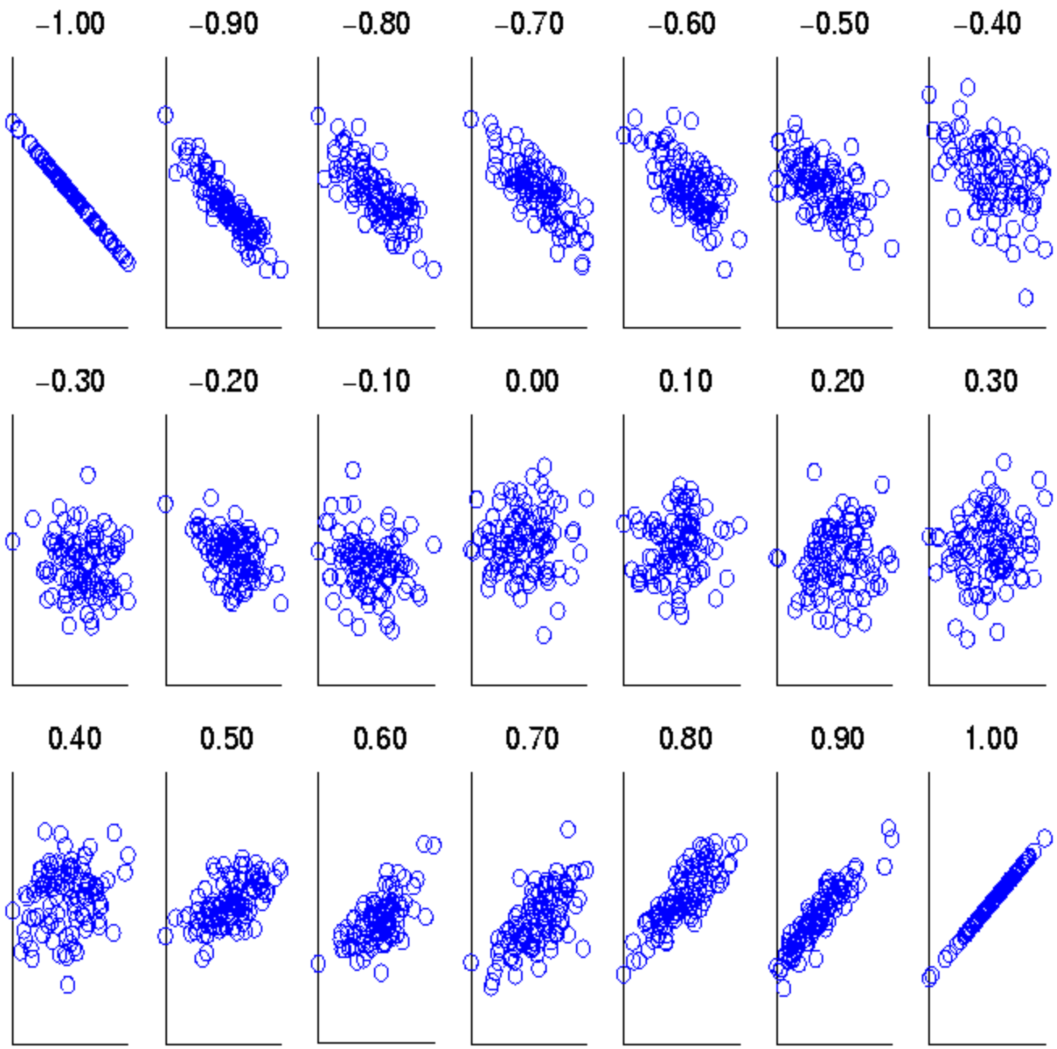
- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

General Approach for Combining Similarities

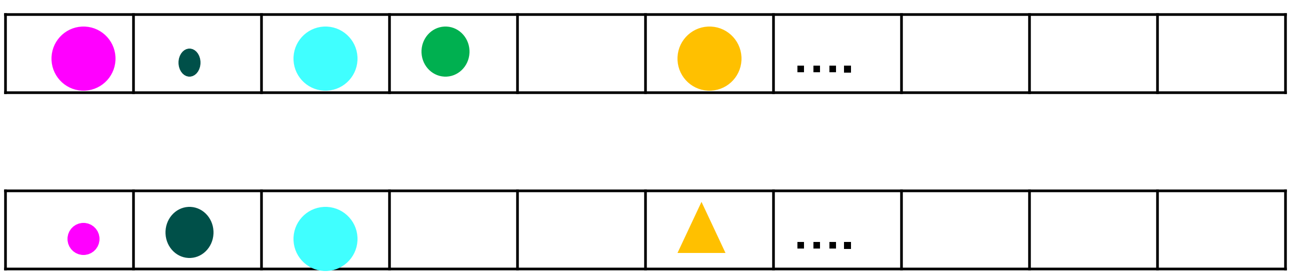
- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$



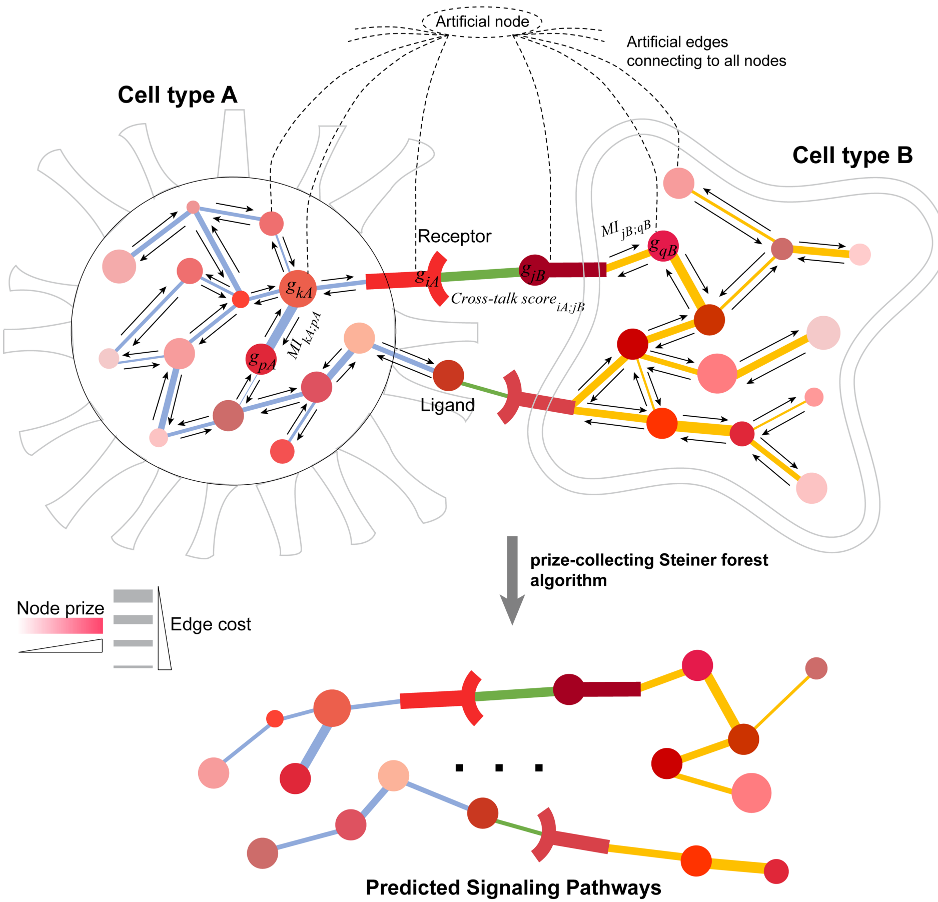
Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

相似性应用：CytoTalk算法解析细胞通讯

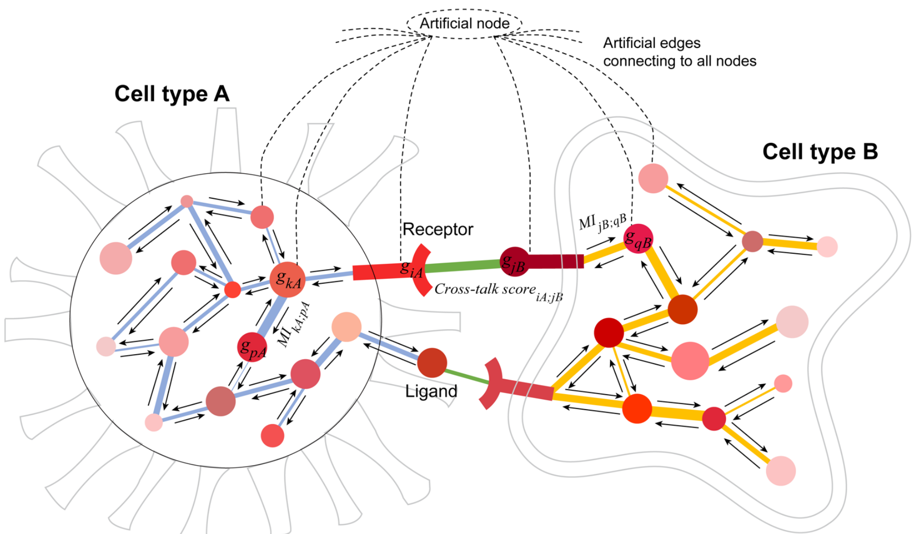


Mathematical modeling:
Formulate the prediction of cell type-specific signaling network as a prize-collecting Steiner forest problem.

Key question:
How to define node prize and edge cost in the input cross-cell-type gene network?

Hu Y, Gao L, et al. *Science Advances*, 2021

相似性应用：CytoTalk算法解析细胞通讯



Node prize:

$$Prize_{iA} = Relevance_{iA} \times PEM_{iA}$$

(1) Cell-type-specificity (preferential expression measure)

$$PEM_{iA} = \log_{10}(Expr_{iA} / e_{iA})$$

$$e_{iA} = \sum_{m=1}^{M} Expr_{im} \cdot \frac{s_{*A}}{\sum_{m=1}^{M} s_{*m}}$$

(2) Closeness to the ligand or receptor genes in the network

$$\mathbf{Relevance}_{A}^{t} = \alpha \mathbf{W}' \mathbf{Relevance}^{t-1} + (1-\alpha)\mathbf{Relevance}^{0}$$

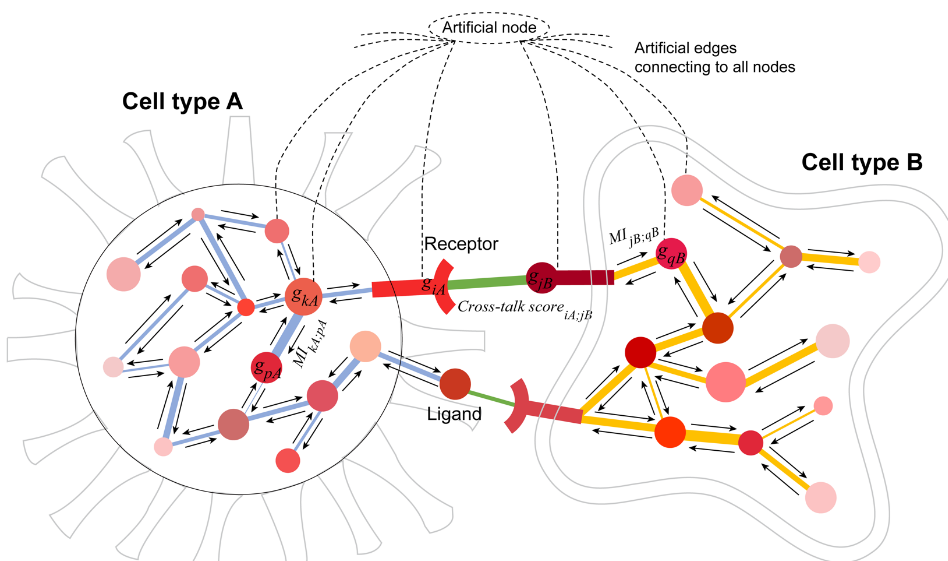
相似性应用：CytoTalk算法解析细胞通讯

Edge cost：

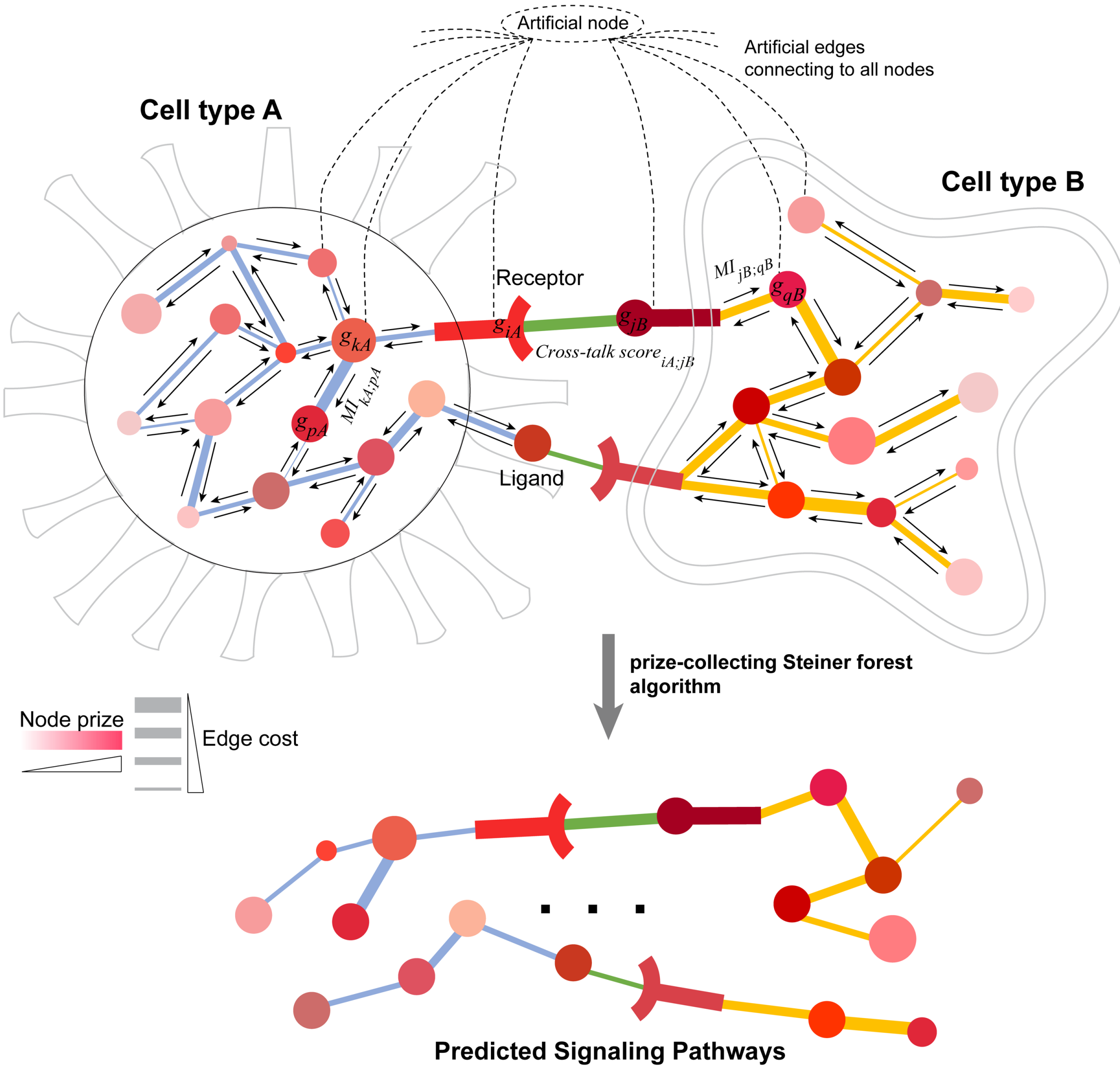
$$Crosstalk\ score_{iA,jB} = Norm\left(Expression\ score_{iA,jB}\right) \times Norm\left(Non-self-talk\ score_{iA,jB}\right)$$

$$Expression\ score_{iA,jB} = \left(PEM_{iA} + PEM_{jB}\right)/2$$

$$Non-self-talk\ score_{iA,jB} = \left[\left(-log_{10}\frac{MI_{iA;jA}}{\min\{H_{iA}, H_{jA}\}}\right) + \left(-log_{10}\frac{MI_{iB;jB}}{\min\{H_{iB}, H_{jB}\}}\right)\right]/2$$



相似性应用：CytoTalk算法解析细胞通讯



Objective function:

$$\min_{F} c(F) + \beta \times p(\bar{F}) + \omega \times k$$

感谢各位同学！

下次课再见！