A Very Brief Glimps On Complex Networks

陈思宇 sychen@zju.edu.cn

2017 年 5 月 3 日

摘要

This article introduces the concepts, methods, and some newest research of complex networks. Since the area of complex networks covers massive contents and topics, in this article, we only promise the delivery of a very short yet brief glimps on Complex Networks.

目录

1 Why complex networks?

The disscussion about complex networks often comes to the description and definition of it. However, respecting the fact that the definition of complex networks is pretty general, we shall firstly take a look at some concepts.

- Community evolution
- Overlapping communities
- Directed networks
- Community characterization interpretation
- Modularities

Of course, those are all bullshits and very hard to understand neither separately nor integerally.

Because of the complexity of network structure and representation, like the relationship between human beings, the description of networks' characteristics are made from a great deal of aspects. That partially explains why there are so many annoying sophisticated concepts.

While the massive concepts obscure the fundaments of complex networks, we can always find it's true nature through it's purpose. On end, the purposes explain the methologies.

The purposes of complex networks research is mainly focused on the following two aspects.

- Finding unifying principles.
- Exploring the dynamics.

2 Is complext-networks' research useful?

Yes, it is useful in our daliy life. For example, like the following applications.

- finding prevailing products and set promoting policy / strategies.
- finding key users. key users are those who have considerable influence over other users on purchase decisions.

- pinpointing hot topics and delivering ads.
- investigating social networks' following relationships to provide sensible search sortings.
- clustering genes in biological research, to find the closeness-relations between different gene serials and RNA clips in regard of their functionality and structure similarity.

3 What is Complex Networks?

As mentioned before, the definition of Complex networks is hard to properly summerized. But, extracting those most essential elements, we can refer to the definition given by QIAN Xue Sen.

- self-organizing
- sefl-similar
- attractor (simple,singular)
- scale-free (power-law)
- small-world

Self-organizing describes the empirical phenomenon that complex structures are actually generated by individuals with various characteristics.

Self similarity means fractal property, as the picture shown below: The top level view of the networks is similar to any sub-view into it's arbitrary branch.

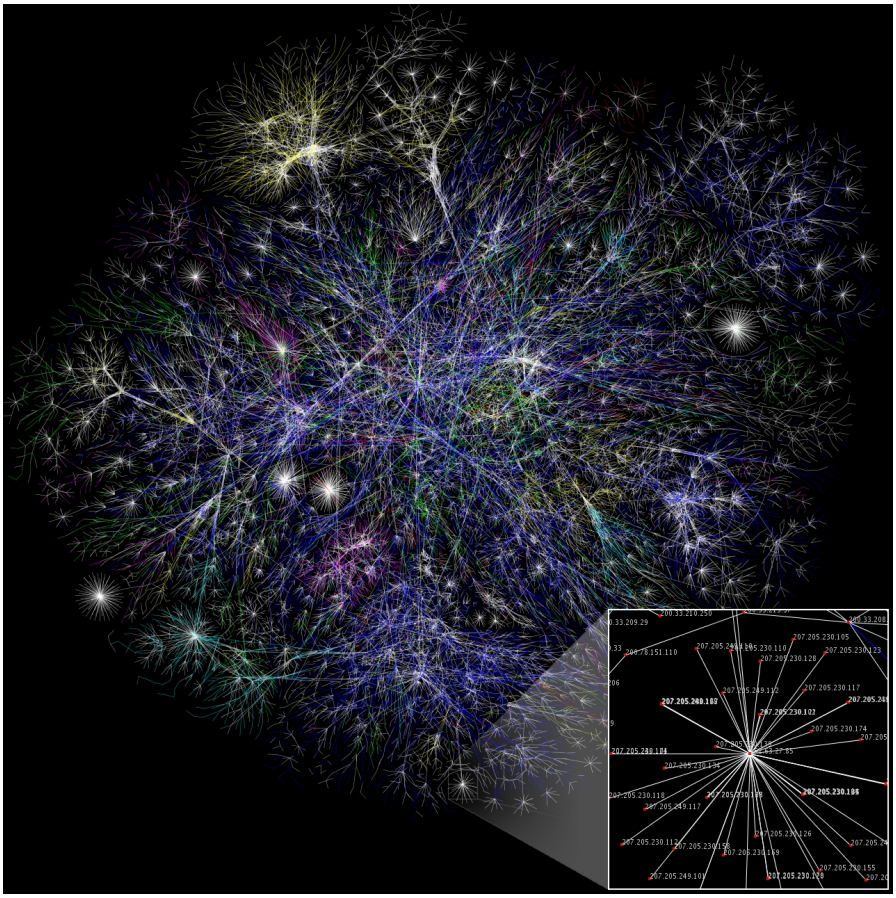

图 1

Attractor is a state of converge of some complex system. If the state of the system converges to some stable ones, then those states are called simple-attractor. In the opposite case, they are called singular-attractors.

Scale-free is also refered as Power or Long-tail principle, which comes from the statistics of english words' frequency.

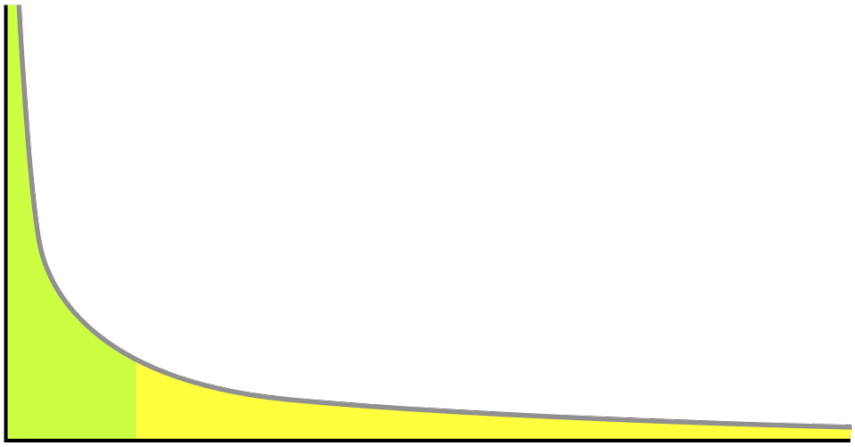

图 2

From the chart we see that the probability of a word being preferntially used is of inverse exponential relationship to it's frequency of appearance. This is called Long-tail phenomenon.

Later, the concepts purposed like Degree, Closeness(∞), and Harmonic Centrality all show explicit Power-law characteristics.

Small-world property demonstrates that the any individual in a complex network tend to have relatively short paths that connect to anyone else.

Here we shall mention the small world experiment carried out by Milgram in 1967. He randomly send letters to others in another city requesting them to find a way to send those letters back to them as the graph shown below.

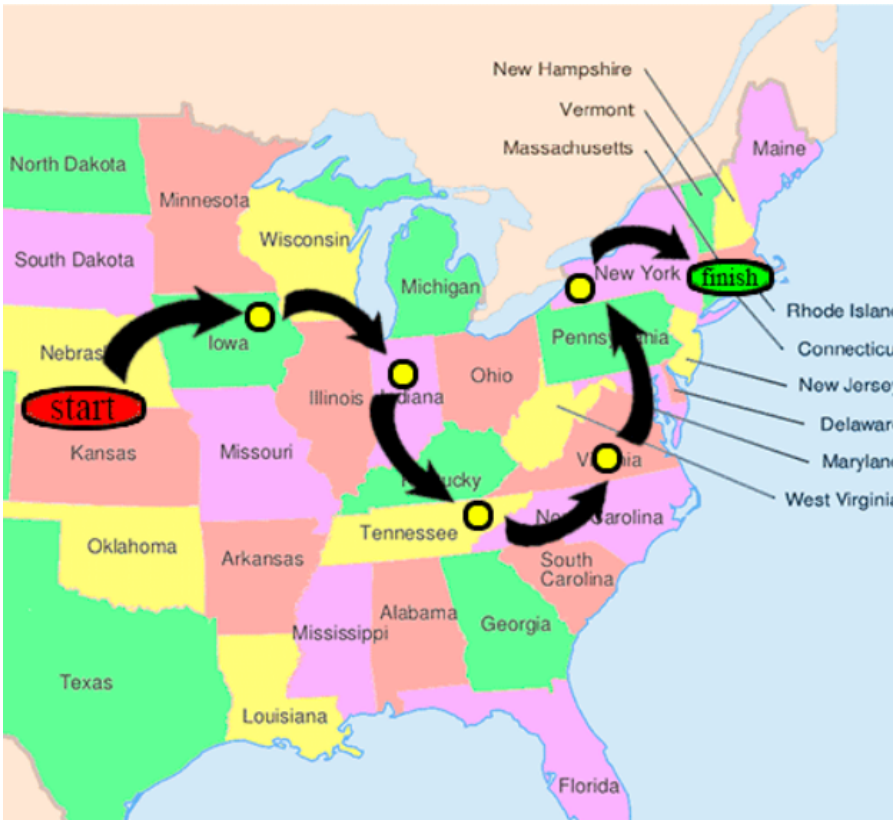

图 3

The results are,

- 232 letters in 296 are lost.
- 64 letters are returned.
- average hops is between 5.5 and 6.

4 Some Characteristics

- Assortative Mixing: $E(i,j), \Delta(node_i, node_j)$ is small.
- Density $\frac{\text{avg deg}}{\text{complete deg}}$ smells cluster.
- Pearson corl-coeff: $cov(x,y) = E[(x - E[x])(y - E[y])]$, $\frac{cov(x,y)}{\sigma_x \sigma_y}$
- Scalar attributes tends to be correlated.

- Probablistic properties: hidden vars behined time span.(Gaming)

Different complex networks own different properties. Some networks have intrinsic aggregations while some networks don't.

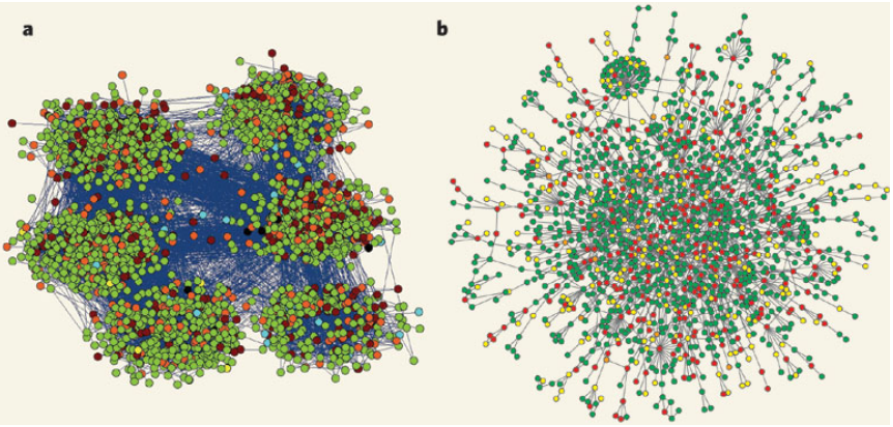

图 4

a: school communities; b: proteins in brewer's yeast.

Now we can know clearly and easily about the following concepts.

- Clique, which is indeed complete subgraphs.
- Density, which is defined as $\frac{\text{avg deg}}{\text{complete deg}}$
- Modularity: used to describe the extent of module-divided formation of networks. Defined as: measurement of extra edges aside of random connection.

5 Algorithms

Algorithms concerning about complex networks can be classified as four classes.

- Heirachical
- Information theory
- Graph theory
- Other stupid fantacies

Nevertheless, the algorithms are not restricted to any one specific class nor type. Instead, they are often combined methods involving many different ideas.

5.1 Erdos-Renyi probability based graph model

Erdos-Renyi probability based graph model was proposed in 1959. The key idea of this algorithm can be stated as 'It is with a probability are any two nodes linked'.

Based on this probablistic view, the existence and creation-process of edges, degrees, clusters and communities can be systematically explained.

(m is the number of edges, k is degree)

- $Pr(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$
- $\mathbb{E}[m] = \binom{n}{2} p, \mathbb{E}[k] = \frac{2}{n}\binom{n}{2} p$
- $Pr(k) = \binom{n-1}{k} p^k (1-p)^{(n-1-k)}$

  $Pr(k) \simeq \frac{c^k}{k!} e^c$ (Poisson)

Under Erdos-Renyi's assumption, the Phase-transition between "having at least on big community" and "having no community" can be explained and resolve to the edge-linking probability. (S represents the proportion of the entire networs' nodes that are in the big community)

- $Pr(\neg e_{i,j}) = 1 - p$
- $Pr(e_{i,j} \wedge \neg e_{j,compo}) = pu$
- $u = (1 - p + pu)^{n-1} = [1 - \frac{c(1-u)}{n-1}]^{n-1}$
- $\lim_{n\to\infty} u = e^{-c(1-u)}$, $Pr(e_{i,comp}) = 1 - u = S$
- $\lim_{n\to\infty} S = 1 - e^{-cS}$

We can know that for this network to have a big community, c should at least be bigger than or equal to 1.

5.2 Girvan - Newman algorithm

Girvan - Newman algorithm was proposed in 2004. This algorithm is sometimes wrongly called NG.

Some preceding concepts:

- Vertex bewteeneess: The number of shortest path linking any two nodes which go through this vertex.
- Edge betweenness: as-is above.
- Pareto principle: 80/20 rule: 20% of the nodes owns 80% of the total degrees.

GN algorithm is a breadth first heirachical method.

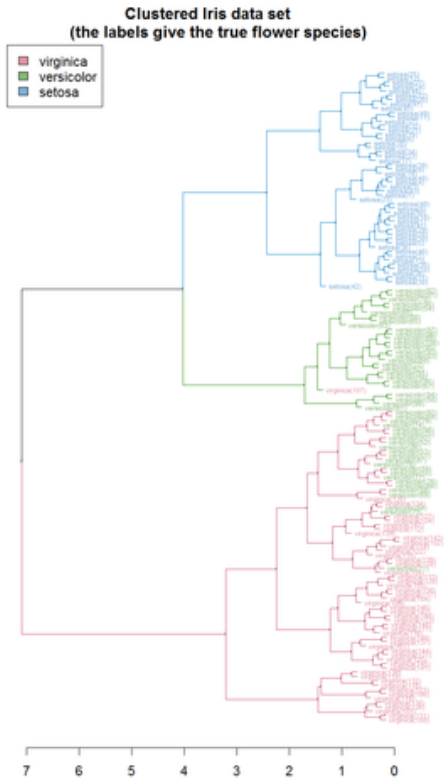

图 5

It continuously remove the edge with highest edge-betweeness to bi-partition the network.

5.3 Latent Semantic Analysis

Latent Semantic Analysis was proposed in 1988. This is a method to analyze the semantic classifications of words and sentences.

(D means documents, W words, Z topics)

$$D = \{d_1, d_2, .., d_N\} \tag{1}$$

$$W = \{w_1, w_2, .., w_M\} \tag{2}$$

$$Z = \{z_1, z_2, .., z_K\} \tag{3}$$

$$A_{coappearance} \text{ is } N \times M \tag{4}$$

$$A = U_{N\times r}\Sigma_{r\times r}V^T_{r\times M} \text{ (SVD/NMF)} \tag{5}$$

Z can be found by applying SVD or NMF technique.

5.4 Probablistic Latent Semantic Analysis

Probablistic Latent Semantic Analysis was proposed in 2000 by Thomas Hofmann. The idea is similar to LSA, but EM is used to determine best topic mappings.

$$P(d_i, w_j) = P(d_i)P(w_j|d_i), \tag{6}$$

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{7}$$

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log P(d_i, w_i) \tag{8}$$

EM is dedicated to find best conditional probabiliy about z topics.

$$P(z_k|d_i, w_j) = \frac{P(z_k)P(d_i|z_k)P(w_j|z_k)}{P(d_i, w_j)} \tag{9}$$

$$\mathrm{E}[\mathcal{L}] = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\sum_{k=1}^{K}\cdots \tag{10}$$

$$P(z_k|d_i, w_j)\log P(w_j|z_k)P(z_k|d_i) \tag{11}$$

But, in order to make things clear, we should take a look back on the EM-method. Furthermore, if we want to know EM better, Lagrange's duality transformation is not to be passed. Hence, we will briefly take a recall on dual problems first.

5.5 Dual Problem

Dual problem is aimed to transform inequity-constrained maximization problem to an equity-constrained minimization one.

- $f(x), s.t.g(x) \leq 0$
- $L(x, \lambda) = f(x) + \lambda g(x), \lambda \geq 0$
- $argmax_\lambda L = f =$ initial problem
- we define original problem as :
  $argmin_x argmax_\lambda L(x, \lambda) s.t. \lambda \geq 0, g(x) \leq 0$
- $L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$
- Karush-Kuhn-Tucker conditions

To obtain the best solution, KKT conditions are deduced.

- $\lambda^* g(x^*) = 0$
- $\frac{\partial L}{\partial x} = 0$
- $\lambda \geq 0, g(x) \leq 0$

KKT is necessary and sufficient condition.

- $\max_\lambda \min_x L(x, \lambda), s.t. \lambda \geq 0, g(x) \leq 0$
- KKT $\rightarrow x = h(\lambda)$
- $\max_\lambda L(h(\lambda), \lambda) s.t. \lambda \geq 0, S(\lambda) = 0$

5.6 Expectation Maximization

Having recalled dual problems, now we take a short tour through Expectation Maximization.

- $f(x; \theta) \rightarrow f(x, y; \theta)$
- randomize θ, y

- estimate $Q = q(y, \theta)$ (E-step)
- $\frac{\partial - L(Q)}{\partial \theta} = 0$,update θ (M-step)
- repeat until converge.

The convergence is guaranteed by the following deduction.

- Jensen's inequality $f(\sum \lambda x) \leq \sum \lambda f(x)$
- $\sum \log \sum p \geq \sum \sum Q \log \frac{p}{Q} = l(\theta)$
- $l(\theta)$ is the lower bound.
- $f(E_z[\frac{p(x,z;\theta)}{Q(z)}]) \geq E_z[f(\frac{p(x,z;\theta)}{Q(z)})]$
- $\frac{p}{Q} = Const$ s.t. $\sum_z Q(z) = 1$
- $Q(z) = \frac{p(x,z)}{\sum p(x)} = p(z|x;\theta)$

  Now we can happily use EM to perform optimization solvings.

5.7 Ball,Karrer,Newman algorithm

Ball,Karrer,Newman algorithm is called BKN for short, and is able to detect overlappings.

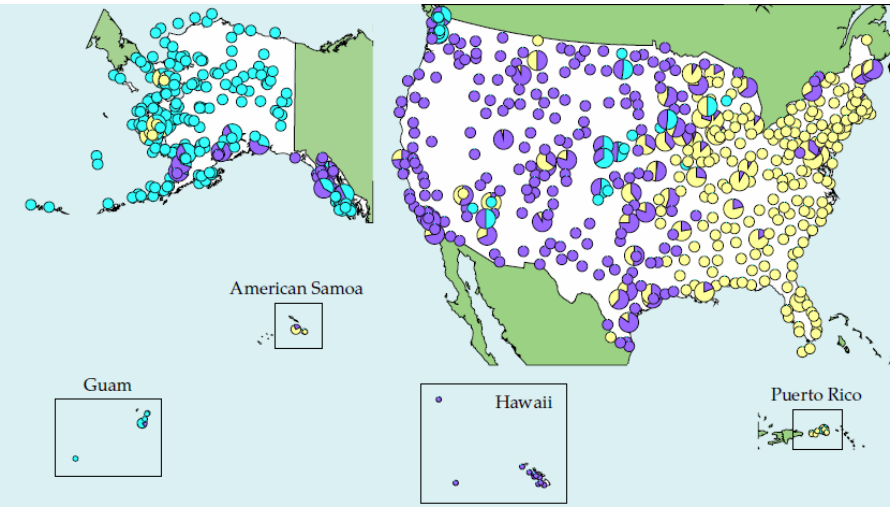

图 6

the figure

5.8 Latent community discovery network

- how to get a list of core actors?
- $\sum_z P(z|d) = 1$,(mixing)
- $\sum_a P(a|z) = 1$ (topics are prevailing)
- $L = \sum_d \sum_a C(a,d) \log \sum_z P(a|z)P(z|d)$
- $R = \sum_{a1,a2} \sum_z \|p(z|a_1) - p(z|a_2)\|^2$ what?

Problem definition and solving.

- $Objective = \alpha(-L) + (1-\alpha)R$
- EM again.

Results on DBLP co-authorship. (AI DB DP GV NC)

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.276	0.238	0.265	19	243	199	244	256	299
k-means	0.257	0.978	0.406	569	2	19	1218	1	1
NG	Unavailable since this network is not fully interconnected								
BKN	0.156	0.306	0.206	799	852	100	80	126	83
LCDN	0.456	0.434	0.445	114	136	370	108	251	376

Table 1: Algorithm performance comparison on the DBLP co-authorship network

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.309	0.245	0.273	18	221	252	242	213	294
k-means	0.258	0.979	0.409	432	1199	19	1	1	2
NG	0.253	0.988	0.403	8625	1217	1	1	1	2
BKN	0.287	0.480	0.359	755	764	102	104	125	127
LCDN	0.501	0.465	0.483	267	394	298	171	92	267

Table 2: Comparison on the fully interconnected DBLP co-authorship network

图 7

Then the experiment was carried out on WEIBO network. 1. Entertainment244 2. Leisure, 333 3. Finance, 297 4. Culture, 185 5. Media, 163

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.627	0.567	0.596	1098	176	200	235	299	271
k-means	0.227	0.715	0.345	**558**	77	108	2	992	2
NG	0.201	**0.873**	0.326	85084	11	1124	21	10	15
BKN	0.528	0.478	0.502	4164	184	270	227	304	196
LCDN	**0.682**	0.611	**0.645**	2067	**282**	**185**	**221**	**277**	**216**

Table 3: Algorithm performance comparison on the WEIBO network

图 8

The results is pretty good. However we should notice that, if you torture the data long enough, data will confess.

5.9 Page rank

Page rank was firstly applied in 1998. The algorithm is a fairly simple one but really works.

- incoming-links: other $\rightarrow$ this site.
- Pagerank(site) $PR(site)$
- $PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \cdots\right) d + \frac{1-d}{N}$
- $(1-d)$

Later, some page-rank-like algorithms were also proposed like:

- Idealrank 2009: local pagerank. far-away PR-values are simplified and unified.
- $\lim_{\substack{edge \to full \\ V_{local} \to V_{global}}} Idealrank = Pagerank$
- approxrank 2009: $approxrank \sim Idealrank$

5.10 Trust Network

Trust networks are related to the topics listed below.

- e-commerce
- resource sharing

- SQA,SDN
- propagenda
- promotion/ads

An algorithm on trust network was proposed by Zhang shaozhong. et al. 2012

- Interactions matrix A
- Successful Interactions matrix T
- Faliures F
- $Believe(v_i, v_j) = \sum \sum p(v_i, v_j) \log \frac{p(v_i,v_j)}{p(v_i)p(v_j)}$
- v: success or failure.

this method use Mutual information as basic metric.
$I(X;Y) = \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
if X, Y, iid. then $I = 0$
else $I > 0$
To find an optimal trusty path, we can apply the following ways.

- Floyd
- Dijkstra

But the further question is how to find communities? Zhang gave a way by minimizing total edge number while maximizing connectivity.

- interconnectivity in v $C_v = \frac{2t_v}{k_v(k_v-1)}$
- E$[C_v] = C = \frac{\sum_{n=1}^{\infty} C_v}{n}$
- characteristic path length: $\frac{\sum_i \frac{\sum_j MinDist(i,j)}{n-1}}{n}$

The setps are like followings.

- $f = a\,(\sum Believe) + bC|_m$
- cut m weak interactions. (heirachical?)

- NP hard!
- a heuristic algorithm:

1 cut m edges $e_1, .., e_m$ that maximize f.

2 add e'_m that maximize f, if $e'_m = e_m$ over.

3 cut one edge that maximize f, then go to 2.