ASONAM 2010

What Can the Temporal Social Behavior Tell us?
An Estimation of Vertex-Betweenness Using Dynamic Social Information

Jing-Kai Lou[1,2], Shou-de Lin[3], Kuan-Ta Chen[1], Chin-Laung Lei[2]

Institute of Information Science, Academia Sinica[1]
Department of Electrical Engineering, National Taiwan University[2]
Department of Computer Science & Information Science, National Taiwan University[3]

2010, 08/09

A Research Issue

Finding the important (or influential) individuals in the crowd

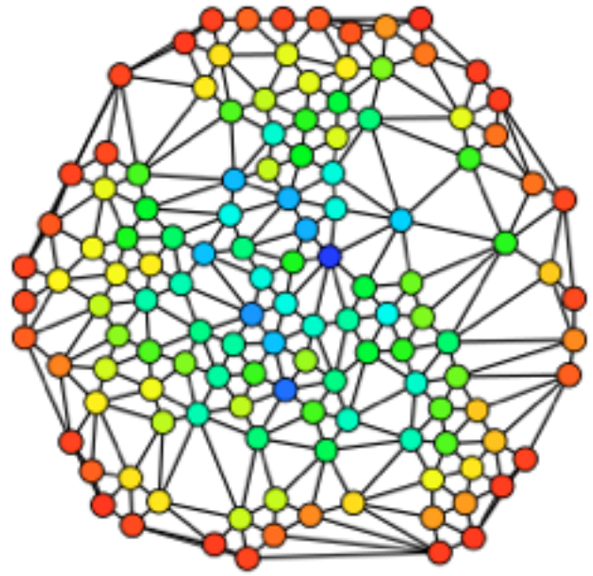
But, how to define the importance?

Centrality in SNA

- To determine the importance of vertices in social networks
- Freeman L.C. [1979]
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality

Betweenness

- Capability to be a communication bridge
- It (from red = 0 to blue = max) shows the vertex-betweenness



$$C_B(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} is the number of shortest paths from s to t,

and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through the vertex v

Time Complexity of Scoring Betweenness

- Calculating the betweenness centrality is time consuming
- It involves finding shortest paths between all pairs

Algorithm Name	Time Complexity
Floyd Warshall [1962]	$O(V^3)$
Johnson [1977]	$O(V^2 \log V + VE)$
Brande [2001]	$O(VE)$
Note: V = the number of vertices; E = the number of edges	

Our Finding

- In some networks, local dynamic (temporal) information of a vertex seems to be relevant to its vertex-betweenness value
- Hypothesis: using the local dynamic information of a vertex, it is possible to approximate its betweenness with lower cost.

Dynamic Social Networks vs. Static Social Networks

	Static	Dynamic
Information Type	NO temporal info.	with temporal info.
Target Issues	• Degree Distribution • Diameter of Network	• Network Growth • Change of Diameter over time
Space Complexity	low	high

Representation of Dynamic Social Networks

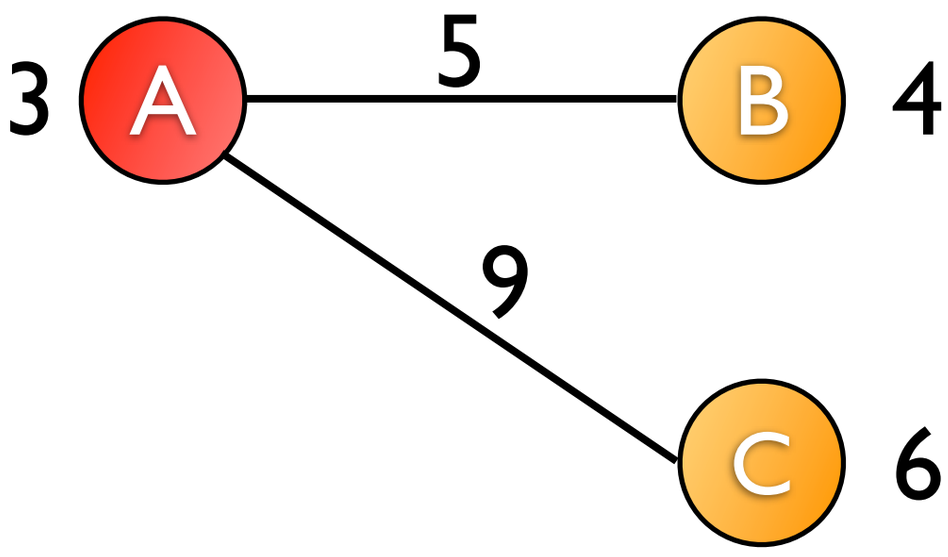
- Notations
  - T_i, the joined timestamp of vertex i
  - C_{ij}, the establish timestamp of the link between vertices i and j
  - The neighborhoods of vertex i
    - N(i) = {j in V | i and j are linked}

Local Dynamic Descriptors for Betweenness Centrality

- **Attraction**
  - The ability to attract newly joined individuals
  - Such vertices might have better chance to play the bridge role in the network.
- **Dynamic Transitivity**
  - The willingness to introduce his/her acquaintances to each other.
  - Such person might have better chance to become the bridge in the network.

Attraction

- Attraction shows how capable a vertex v is in terms of making connection with newly coming individuals.
- a newly coming individual is defined as arriving no earlier than k time units before the link is established
- $N_k(v) = | \{u \text{ in } N(v) \mid C_{uv} \leq t_u + k\} |$



For red vertex,
$N_1 = |\{B\}|$,
$N_3 = |\{B, C\}|$

Thinking Attraction

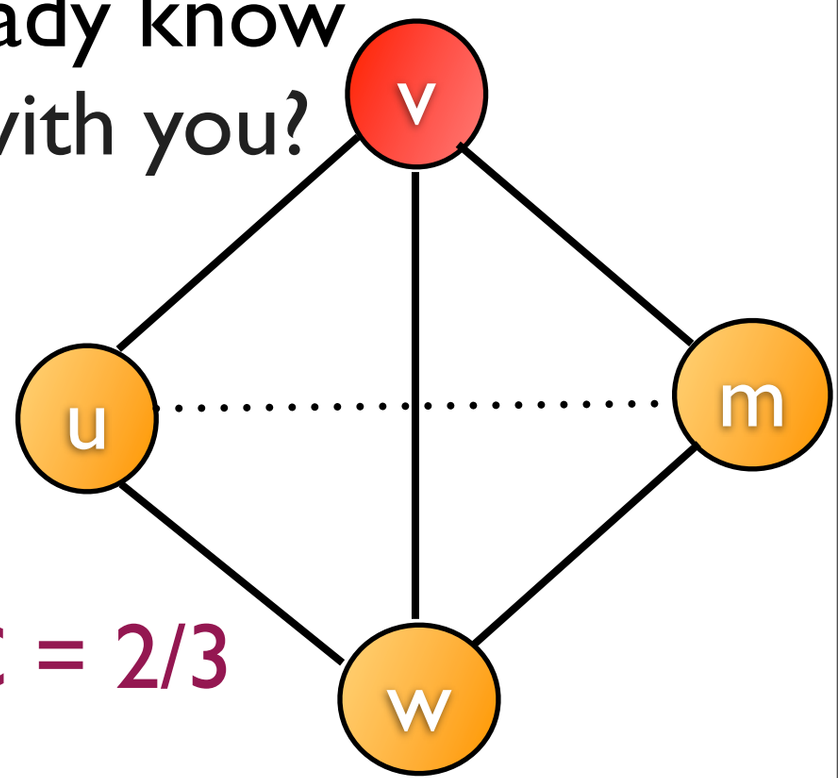
- Attraction equiv. to degree if k is infinity
  - degree captures the connectivity of vertices
  - but it does not capture how well a vertex can attract newly joined vertices

Local Dynamic Descriptors for Betweenness Centrality

- **Attraction**
  - The ability to attract newly joined individuals
  - Such vertices might have better chance to play the bridge role in the network.
- **Dynamic Transitivity**
  - The willingness to introduce his/her acquaintances to each other.
  - Such person might have better chance to become the bridge in the network.

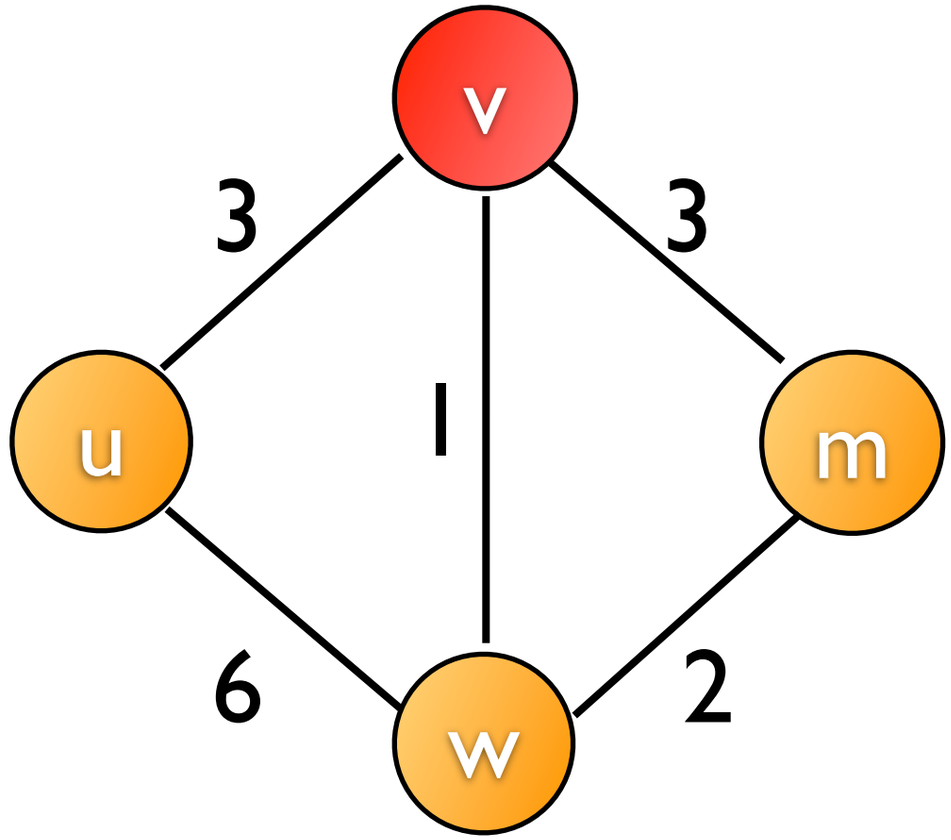
How about Using Clustering Coefficient to Model "*Introduction*"

- Pros: Clustering coefficient can be used to model how the neighbors are connected to each other.
- Cons: Connectivity != Capability of introduction
  - what happens if your two friends already know each other before they are acquaint with you?
  - CC is not the best way to model "the capability of introduction".



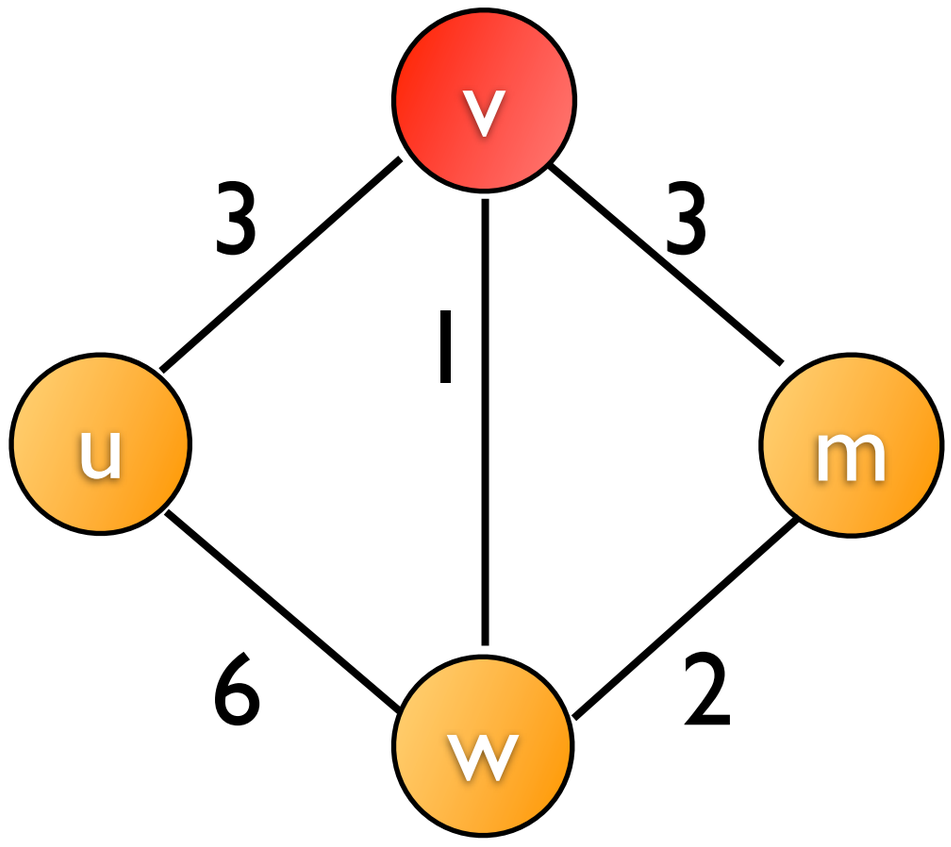
When Temporal Info is Considered

- Definition:
  a pair of vertices (u, w) is said to be introduced by v if and only if $t_{uw} \geq \max(t_{uv}, t_{vw})$
- There is only 1 pair (i.e. (u, w)) introduced by vertex v in the example



Dynamic Transitivity

- Dynamic Transitivity of v, T(v), is the number of pairs introduced by v, divided by the clustering coefficient of v
- $T(v) = | \{(x, y \text{ in } BN(v) \mid C_{xy} >= \max(C_{vx}, C_{vy})\} | / cc(v)$



Pairs Introduced: 1,
cc: 2/3,
Dynamic Transitivity = 3/2

How Time is Reduced based on Our Model

- To collect dynamic transitivity for one vertex, it needs to know the connection among neighbors.
- Time complexity: $O(b^2 V)$
  - b = average degree; V = vertices number
- In 27 Realistic networks [Newman, 2003]
  - Average number of edges (E) = 97,719,006
  - Average degree (b) = 14.79
  - b^2: E = 1: ($6.6 * 10^6$)

Regression Model to Predict BC

- Model: Linear Regression
- Baseline Scheme
  - $BC(v) = w_1*OutDegree + w_2*InDegree + w_3*BdDegree + w4*cc$
- Proposed Scheme
  - $BC(v) = w_1*N_k(v) + w_2*T(v)$,
    where $N_k(V)$ denotes the attraction of v, and T(v) denotes the dynamic transitivity of v

Data Set

- Fairyland Online (MMORPG)
- Collection Duration, 2003 Feb. ~ 2004 Apr.
- Messages over private channels in 8 Realms
  - time-stamp
  - sender ID
  - receiver ID

Tell Channel

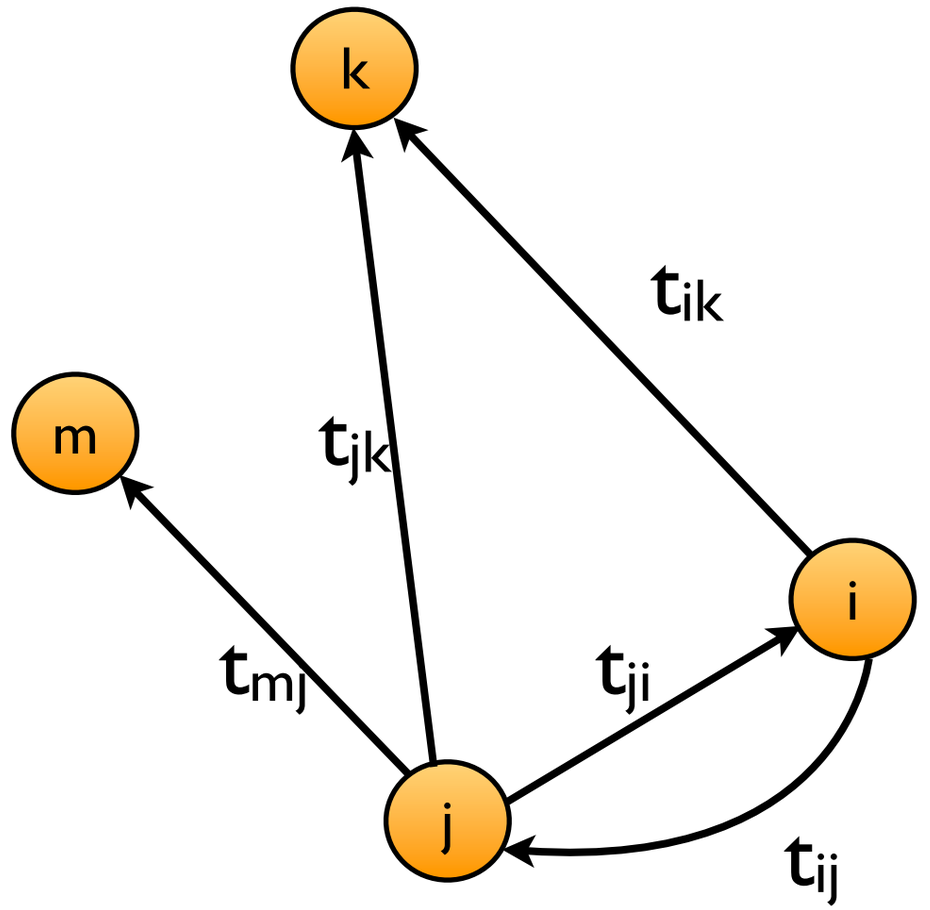
```
Mon May 19 16:02:43 2003 : PetMaster tell vodooxo : damn
Mon May 19 16:03:12 2003 : PetMaster tell vodooxo : run
Mon May 19 16:03:21 2003 : PetMaster tell vodooxo : dont waste time
Mon May 19 16:03:23 2003 : PetMaster tell vodooxo : hahaha
Mon May 19 16:03:41 2003 : PetMaster tell vodooxo : u r just one hardcore killer man
Mon May 19 16:03:51 2003 : vodooxo tell PetMaster : hahah
Mon May 19 16:03:55 2003 : vodooxo tell PetMaster : get exp for wolf
Mon May 19 16:03:58 2003 : PetMaster tell vodooxo : u r soon a WANTED guy from 6fu
Mon May 19 16:04:01 2003 : PetMaster tell vodooxo : lolz
Mon May 19 16:04:08 2003 : vodooxo tell PetMaster : lol
Mon May 19 16:04:15 2003 : vodooxo tell PetMaster : all the owls run from me
Mon May 19 16:04:18 2003 : vodooxo tell PetMaster : the bears run from me
Mon May 19 16:04:21 2003 : PetMaster tell vodooxo : hahahah
Mon May 19 16:04:25 2003 : PetMaster tell vodooxo : ur luck pts
Mon May 19 16:04:28 2003 : vodooxo tell PetMaster : those mud mons hide from me
```

Chat Activity Network

time speaker receiver

Tell Channel

Mon May 19 16:02:43 2003 : PetMaster tell vodooxo : damn
Mon May 19 16:03:12 2003 : PetMaster tell vodooxo : run
Mon May 19 16:03:21 2003 : PetMaster tell vodooxo : dont waste time
Mon May 19 16:03:23 2003 : PetMaster tell vodooxo : hahaha
Mon May 19 16:03:41 2003 : PetMaster tell vodooxo : u r just one hardcore killer man
Mon May 19 16:03:51 2003 : vodooxo tell PetMaster : hahah
Mon May 19 16:03:55 2003 : vodooxo tell PetMaster : get exp for wolf
Mon May 19 16:03:58 2003 : PetMaster tell vodooxo : u r soon a WANTED guy from 6fu
Mon May 19 16:04:01 2003 : PetMaster tell vodooxo : lolz
Mon May 19 16:04:08 2003 : vodooxo tell PetMaster : lol
Mon May 19 16:04:15 2003 : vodooxo tell PetMaster : all the owls run from me
Mon May 19 16:04:18 2003 : vodooxo tell PetMaster : the bears run from me
Mon May 19 16:04:21 2003 : PetMaster tell vodooxo : hahahah
Mon May 19 16:04:25 2003 : PetMaster tell vodooxo : ur luck pts
Mon May 19 16:04:28 2003 : vodooxo tell PetMaster : those mud mons hide from me



Dynamic Social Network,
C_{ij} is defined as first time to establish the bidirectionally link between vertex i and vertex j

Networks Overview

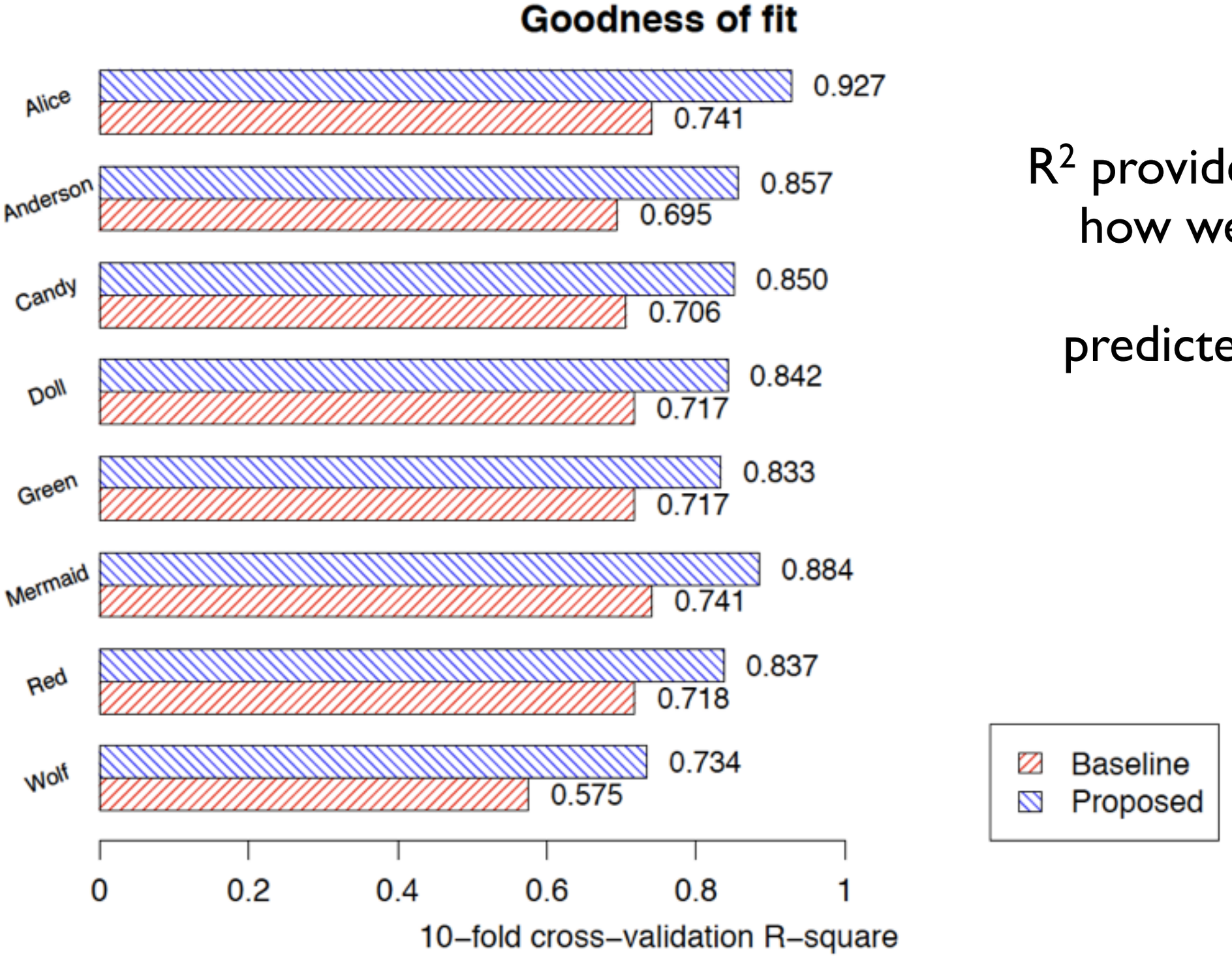
Realm	# node	# edge	avg. deg	avg. cc	gscc ratio	avg. deg in gscc	diameter in gscc	apl in gscc
alice	32,690	445,528	27.2	0.085	66.4	39.0	12	3.70
anderson	101,150	2,363,864	46.7	0.106	79.2	58.1	11	3.74
candy	59,534	1,209,526	40.6	0.094	74.0	53.6	12	3.70
doll	44,891	876,911	39.0	0.107	73.8	51.7	12	3.73
green	88,599	2,073,034	46.8	0.102	78.7	58.5	13	3.73
mermaid	44,656	684,145	30.6	0.081	67.9	43.3	14	3.69
red	60,418	1,239,473	41.0	0.098	76.8	52.3	13	3.74
wolf	54,039	1,233,609	45.6	0.104	75.8	59.0	12	3.72

cc: clustering coefficient; **gscc**: giant strongly connected component; **apl**: average path length

Performance Evaluation

- R-square
- True Values vs. Predicted Values
- Correlation
  - Spearman coefficient
  - Kendall coefficient

Cross-Validation R^2

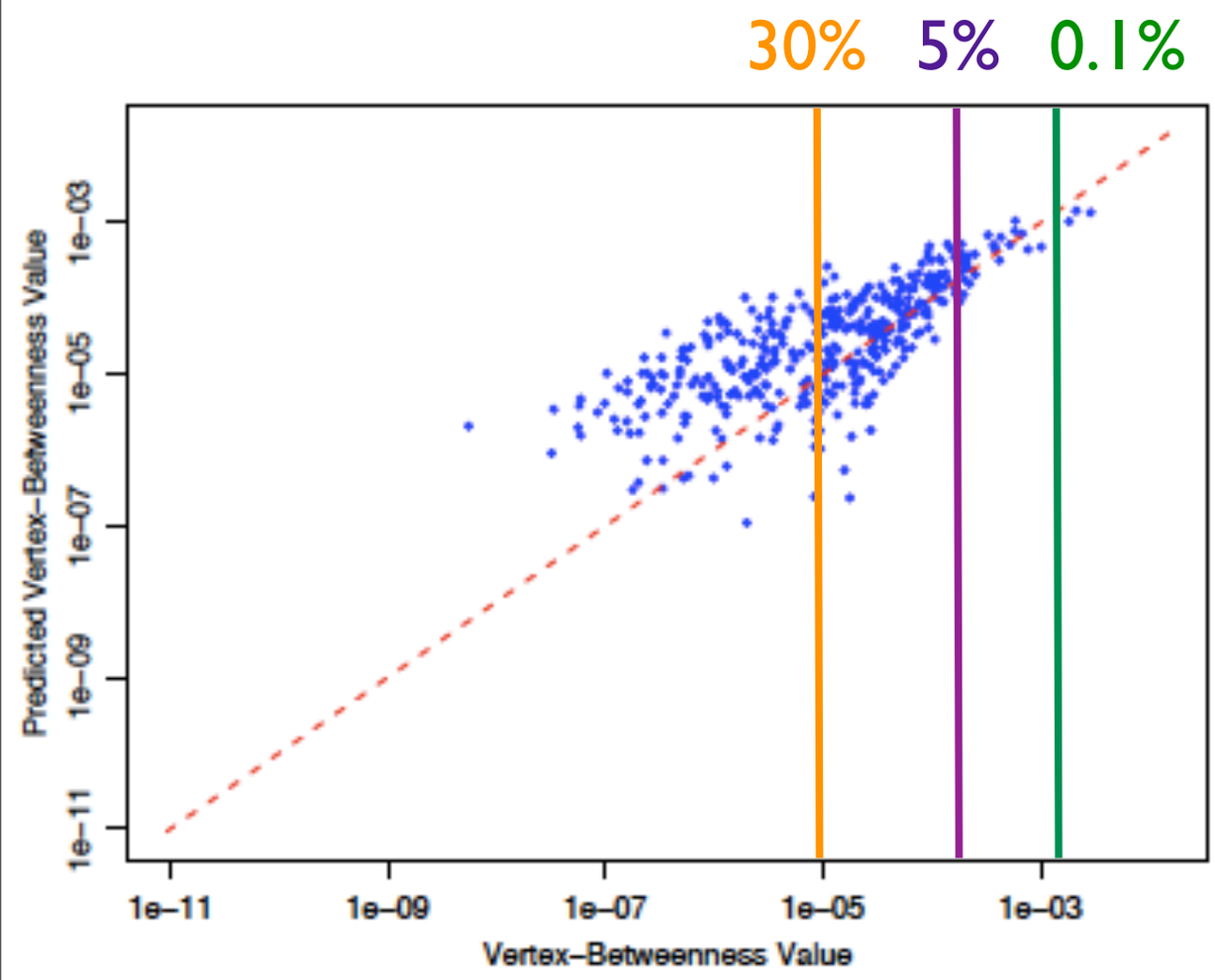


R^2 provides a measure of how well the outcome are likely to predicted by the model

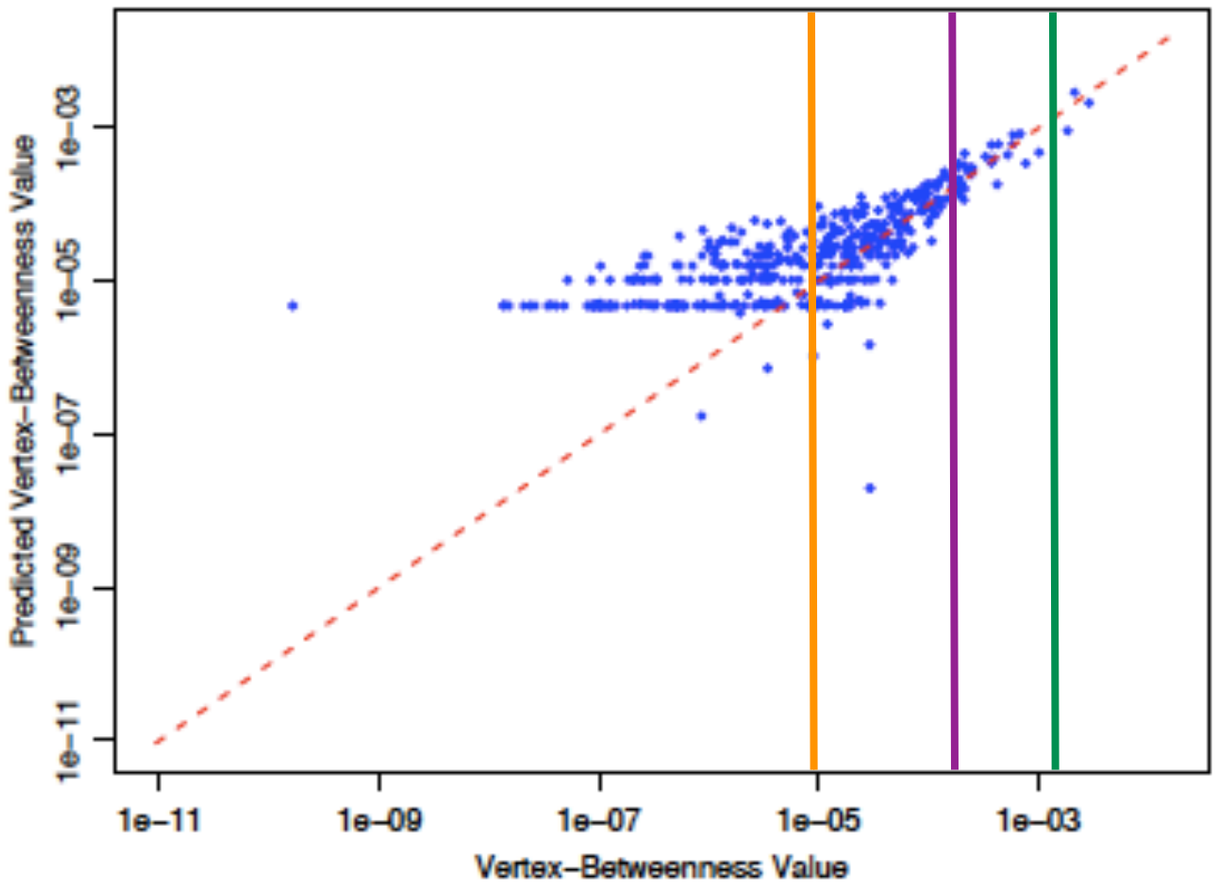
Accuracy: baseline vs model

(x, y) = (truth, prediction)

acc. / threshold	baseline model	proposed model
0.1%	0.693	0.742
1%	0.720	0.793
5%	0.749	0.813
10%	0.735	0.788
20%	0.724	0.767
30%	0.745	0.780

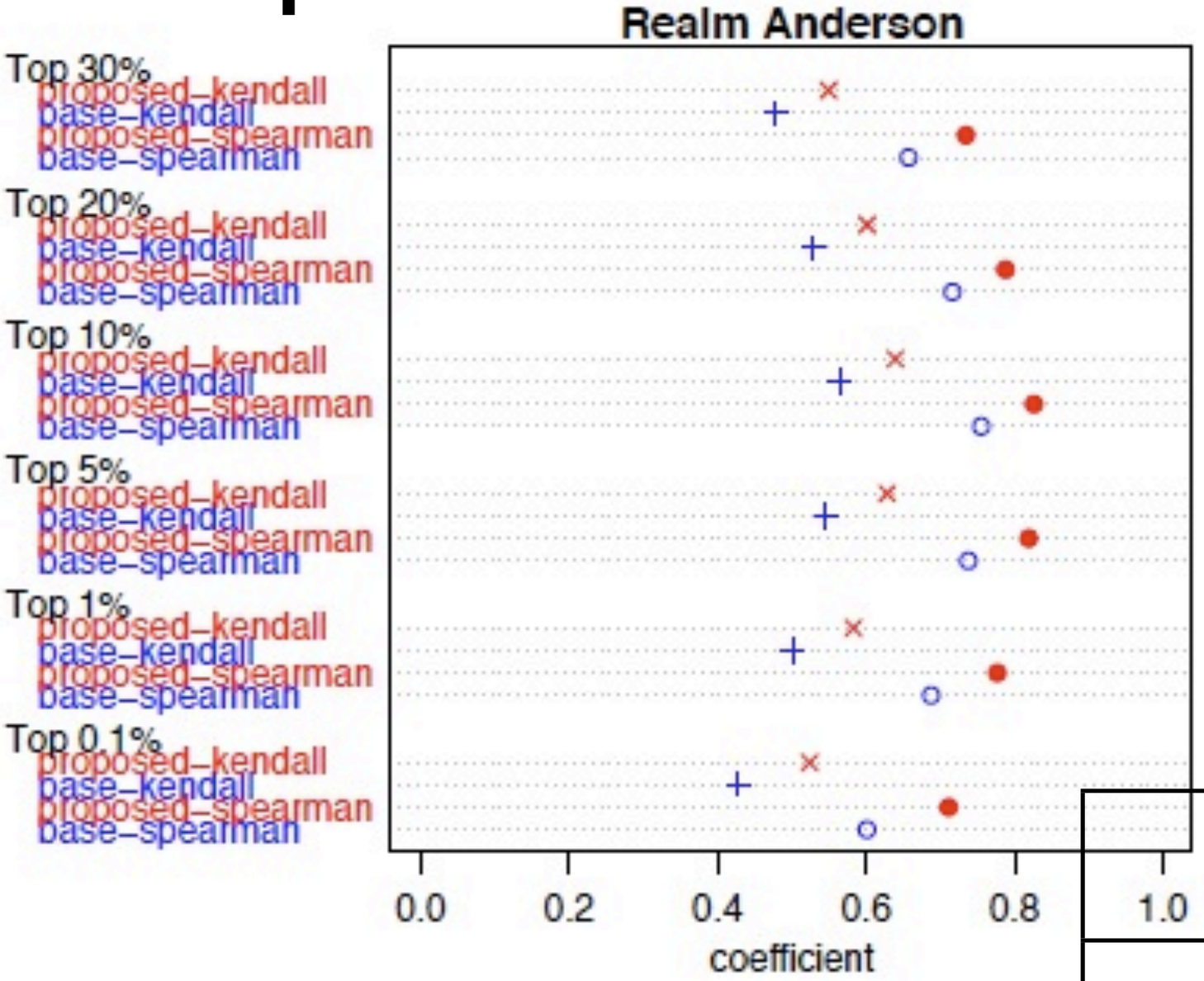


(a) Anderson-base



(b) Anderson-proposed

Top Set Correlation Coefficient



	baseline	proposed
spearman	○	●
kendall	+	x

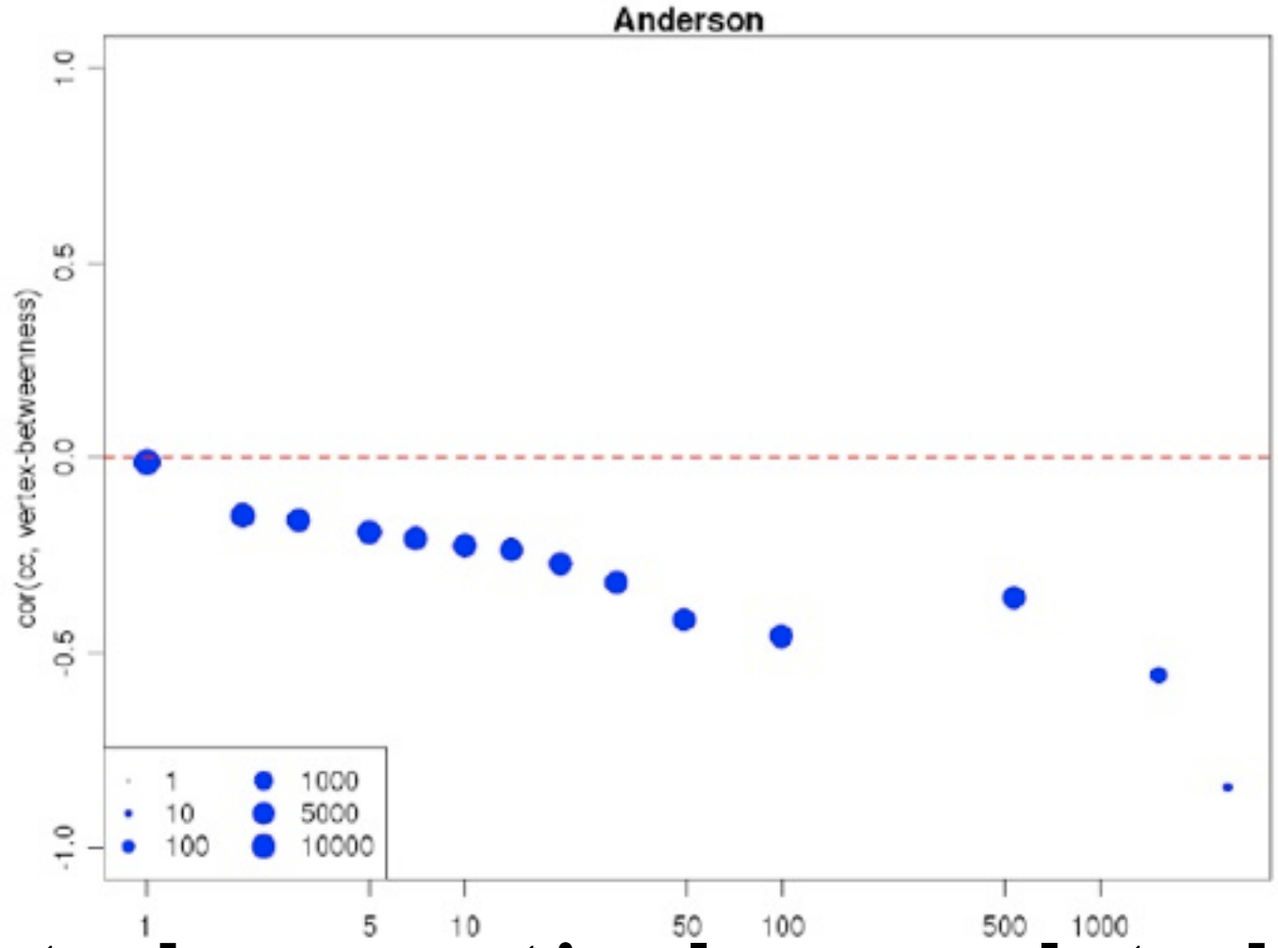
Conclusion

- Observing an interesting phenomenon
  - one can estimate the vertex-betweenness via local dynamic indicators
- Advantages of the estimation
  - fast ($O(VE) \rightarrow O(b^2V)$), rank of estimated values are persisted fairly
- Providing a way to construct dynamic social network for MMORPG, and also analyzing their structural characteristics

Thanks for your attention

Q & A

Clustering coefficient (Transitivity) vs Betweenness



CC seems to be negatively correlated with BC

A vertex X has large CC → X's neighbors have links between each other --> Shortest paths between pairs does not need to go through X --> X's betweenness cannot be too high