“朋友～相似”现象溯源

—利用大数据的分析

朋友间相似的原因？

- 当看到两个关系不错的人在某些特质上相似

相似→朋友？ 朋友→相似？

如果我们能长期观察一群人，看到他们之间关系的演变，以及他们参与的社会活动…

Feedback Effects between Similarity and Social Influence in Online Communities

David Crandall
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
crandall@cs.cornell.edu

Dan Cosley
Dept. of Communication
Cornell University
Ithaca, NY 14853
drc44@cornell.edu

Daniel Huttenlocher
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
dph@cs.cornell.edu

Jon Kleinberg
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
kleinber@cs.cornell.edu

Siddharth Suri
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
ssuri@cs.cornell.edu

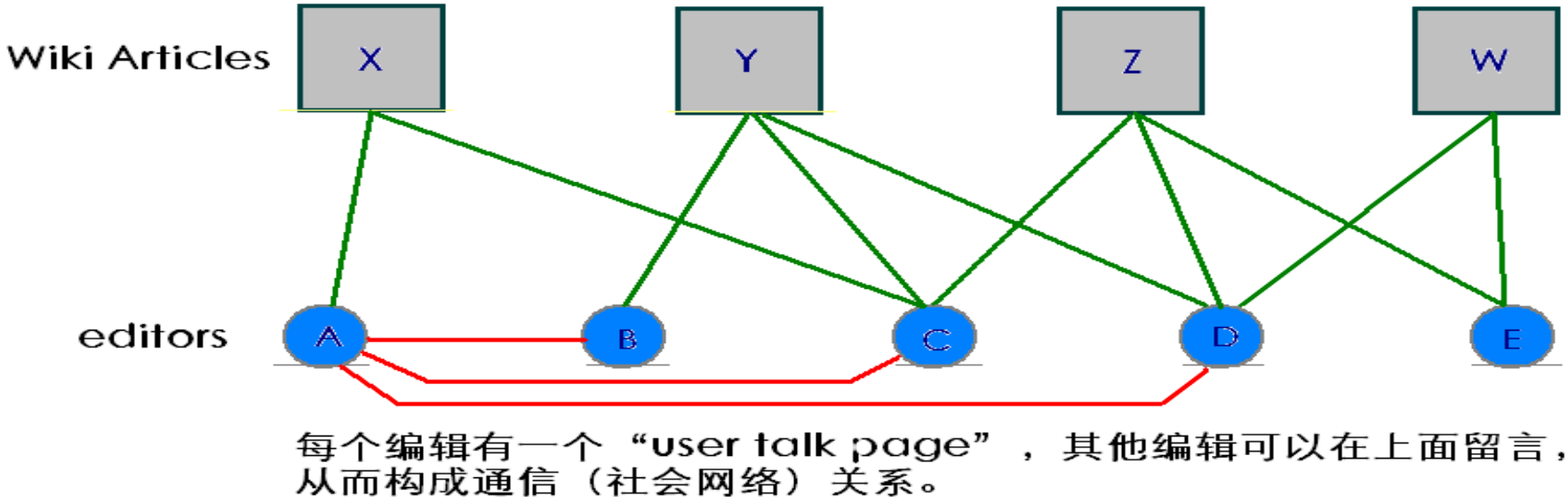
- In Proceedings of 14th ACM SIGKDD, 2008

需要一个数据集

- 反映随时间变化的大规模社会归属网
- 大规模：人多，社交聚点多
- 随时间变化：人和人之间，人和社交聚点之间

英文维基百科数据：50万人，300万文章

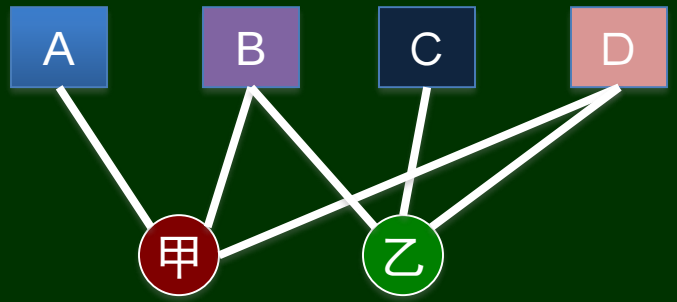
利用在线数据研究同质性现象



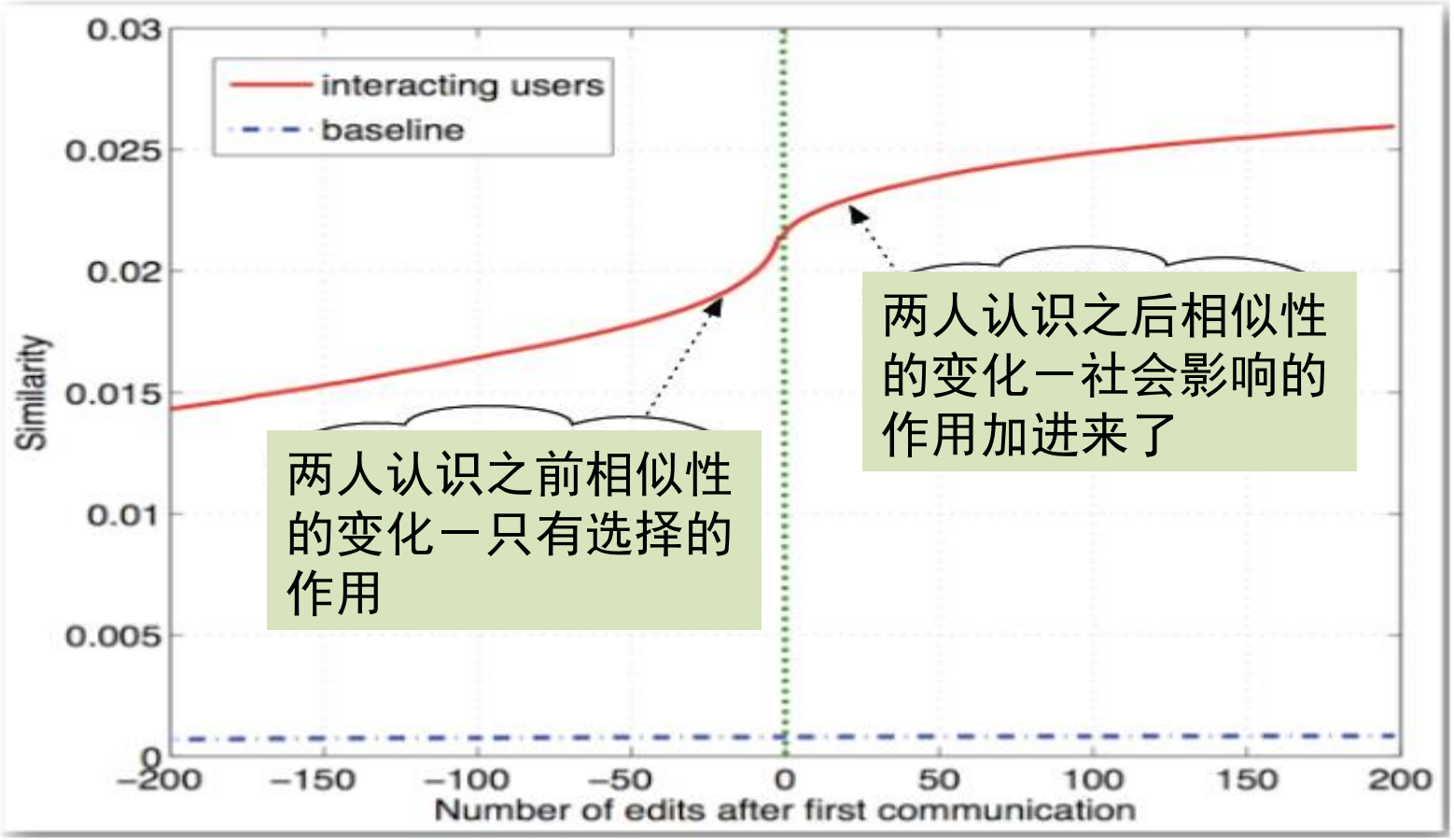
- 两个编辑之间相似性的变化与“自然选择”和“社会影响”的关系
- 没有联系（通信）之前，相似（编辑相同文章）主要因为选择；达到足够相似度时则容易发生联系，然后社会影响开始对相似性提高起作用

两人相似性（度）的测量

$$\text{相似性} = \frac{\text{两人都编辑过的文章数}}{\text{总共编辑过的文章数}}$$



相似性、选择与社会影响



KDD 2008

图 4.13

小结

- 我们展示了一种利用“社会归属网”大数据，剖析同质性现象原因的思路
- 学习的要点是：从问题，到模型（社会归属网），到数据（维基百科），到映射（数据与问题要素的关系）这样一个过程
- 进一步的细节可参见教材64－67页，以及参考文献［122］