UTHealth® Houston
School of Public Health

# Synthesizing Network Science, Graph-Based Deep Learning, and Blockchain in Data Science for HIV Research Addressing Health Inequities and Complex Graph-structured Data

CFAR Symposium on Statistics and Data Science in HIV Providence Center For AIDS Research

Kayo Fujimoto, PhD

June 5-6, 2023

# Outline of this presentation

- Proposed integrated framework: Network science, Graph-structured Data Science Approach, and Blockchain implementation to address HIV Inequities
- Brief Introduction to Social Network Analysis
- Empirical application of social network analysis and graph-based deep learning in HIV research
- Promise of Blockchain implementation to address access to PrEP care continuum for marginalized population

# Key Components of the synthesized framework for addressing HIV inequities

- **Social Network Analysis (SNA)**
  - Studying relationships and understanding social structures and dynamics and their impact on social phenomena
- **Graph-based Deep Learning (GBDL)**
  - Learning and extracting features of complex network structures and individual attributes using deep learning techniques
- **Blockchain concept/technology (Blockchain)**
  - The use of cryptography to secure data, decentralized governance, and distributed digital ledger technology to ensure security, transparency, and immutability in the storage, management, and sharing of data among stakeholders.

# Proposed integrated framework for graph-structured data science approach and blockchain implementation within network science

- Gain a more complete understanding of health inequities from a network perspective
- Uncover hidden structural patterns
- Make accurate predictions
- Ensure data security and privacy
- Actively engage communities in the research process
- Inform targeted interventions and policy development
- ➡ Creating a more inclusive and equitable health landscape.

# Introduction to Social Network Analysis

UTHealth® Houston
School of Public Health

# Network Science and Social Network Analysis

Network Science is defined as the "study of the collection, management, analysis, interpretation, and presentation of relational data."(p. 3 in Brandes, Robins, McCranie, & Wasserman, in Network Science (2013))

Social Network Analysis provides a common set of methodologies and analytical tools to visualize, measure, and make a statistical inference for various network characteristics (Wasserman & Faust, 1994).

Sources:
Brandes U, Robins G, McCranie A, Wasserman S. What is network science? (2013). *Network science*. *1*(1):1-5.
Wasserman S, Faust K. (1994). *Social network analysis: Methods and Applications*.

# Utilities of Social Network Analysis

Social network analysis permits us to:

Understand patterns of relationships and network structures

- Visualize patterns
- Describe, measure, and model structural features
- Make statistical inferences about how relations are patterned
  - ➡ Exponential Random Graph Models (ERGMs)

# Handling social network data

- Social network data:
    - Recorded by edge list/dyad format or matrix representation
    - Visualized by a graph
    - Structural characteristics are computed by matrix algebra or computer algorithms
        - Network/Sub-network-level: density, network size
        - Node-level: centrality, isolates

# Application of Social Network Method to HIV/STI Research

UTHealth® Houston
School of Public Health

# YMAP: Young Men's Affiliation Project of HIV Risk and Prevention Venue

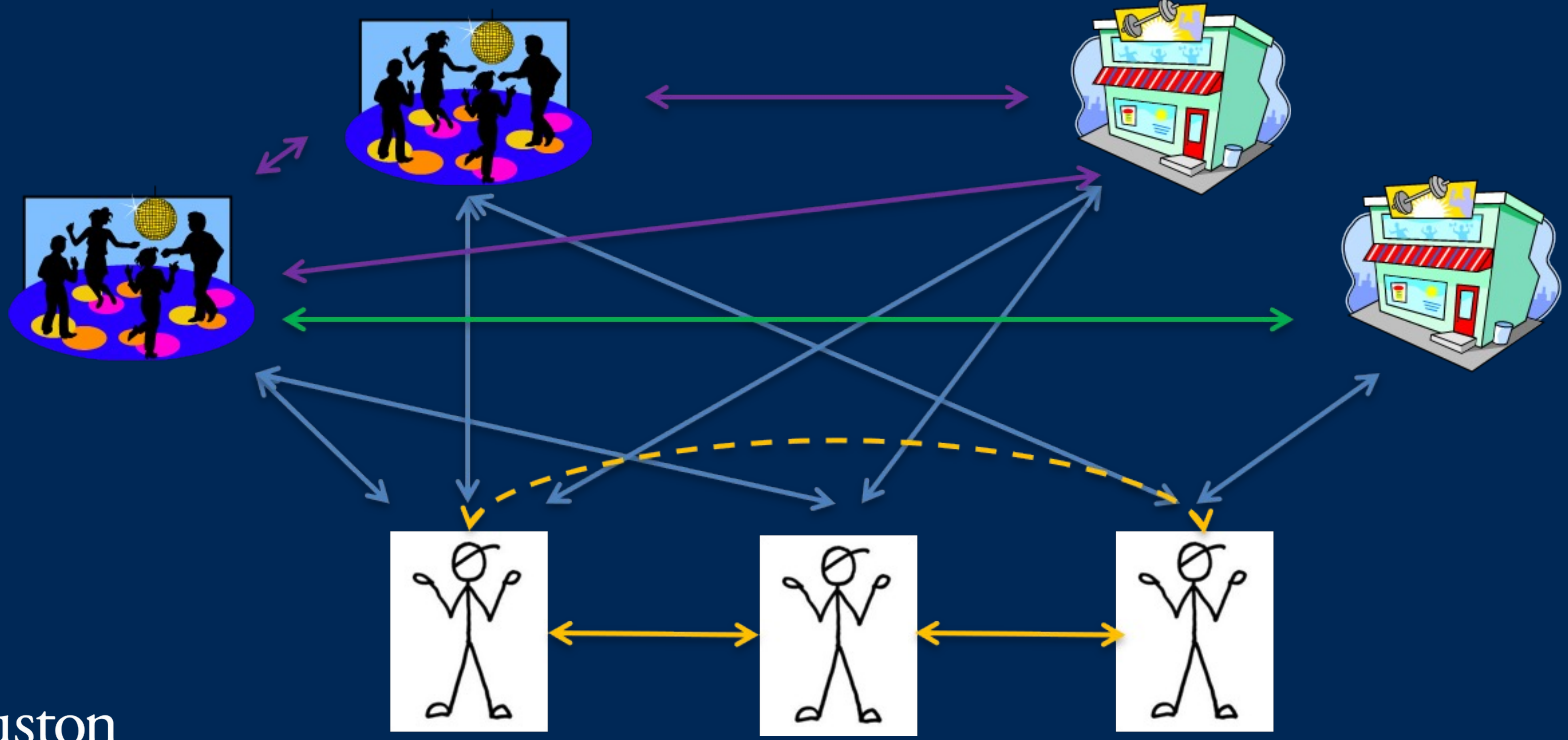
NIH/NIMH 1R01MH100021

(PI: Fujimoto, K. & Schneider, J.A.) 2013-2018

UTHealth Houston
School of Public Health

# Description of YMAP

Use social network analysis to examine complex multilevel networks among younger sexual minority men aged 16-29 and venue and health organizations in Houston and Chicago



UTHealth Houston
School of Public Health

# Study design

- At phase I: Sample of venue representative form social& preventive venues
  - 150 venues in Houston & Chicago
    - Wave 1 (2014)
    - Wave 2 (2015)
    - Wave 3 (2016)
- At phase II: RDS Sample (Network-based link-tracing chain referral recruitment method) of 755 sexual minority men in Houston & Chicago
    - Wave 1 (2014-15)
    - Wave 2 (2015-16)

→ The retention rate was 79%

YMAP YOUNG MEN'S AFFILIATION PROJECT

UTHealth Houston School of Public Health

# Handling social network data

- Social network data:
  - Recorded by edge list/dyad format or matrix representation
  - Visualized by a graph
  - Structural characteristics are computed by matrix algebra or computer algorithms
    - Network/Sub-network-level: density, network size
    - Node-level: centrality, isolates

Social Networks 54 (2018) 118–131

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

# Statistical adjustment of network degree in respondent-driven sampling estimators: Venue attendance as a proxy for network size among young MSM

Kayo Fujimoto[a,*,1], Ming Cao[a,1], Lisa M. Kuhns[b,c], Dennis Li[a], John A. Schneider[d]

[a] Center for Health Promotion and Prevention Research, Department of Health Promotion & Behavioral Sciences, Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, 7000 Fannin Street, UCT 2514, Houston, TX 77030-5401, United States

[b] Division of Adolescent Medicine, Ann & Robert H. Lurie Children's Hospital, Chicago, IL, United States

[c] Northwestern University, Feinberg School of Medicine, Department of Pediatrics, 225 E. Chicago Avenue, # 161, Chicago, IL 60611, United States

[d] Department of Medicine and Public Health Sciences and the Chicago Center for HIV Elimination, University of Chicago, 5837 South Maryland Avenue MC 5065, Chicago, IL 60637, United States

A R T I C L E I N F O

*Article history:*

*Keywords:*
Respondent-driven sampling method (RDS)
HIV/STI
Two-mode affiliation networks
Young men who have sex with men (MSM)
sexual and drug use behavior
Sexually transmitted infections
Pre-exposure prophylaxis (PrEP)
Venue affiliation

A B S T R A C T

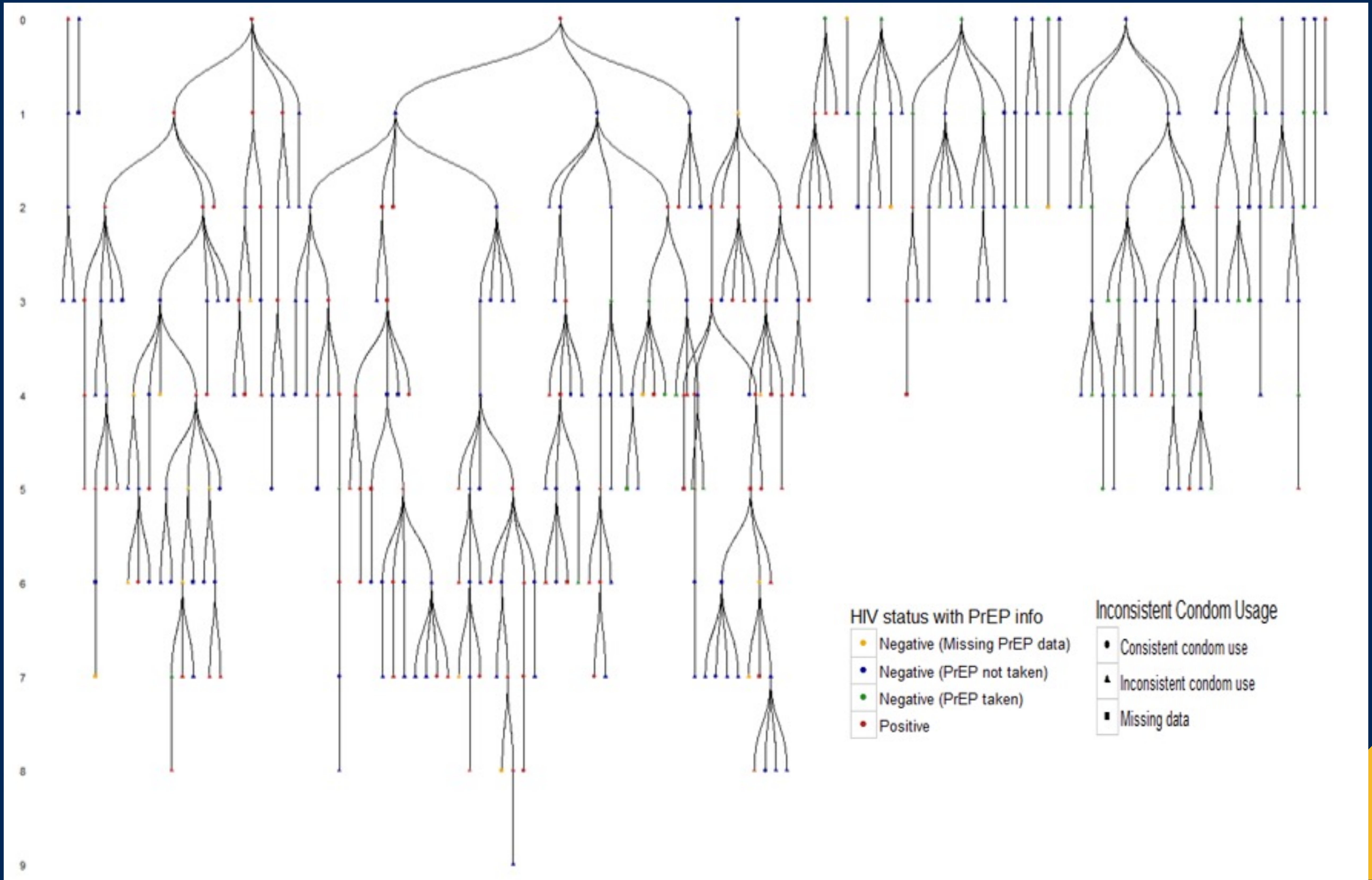
We introduce a new venue-informed network degree measure, which we applied to respondent-driven sampling (RDS) estimators. Using data collected from 746 young MSM in 2014–2016 in Chicago, IL, and Houston, TX, we estimated the population seroprevalence of HIV and syphilis and risk/protective behaviors, using RDS estimates with self-reported network size as a standard degree measure as well as our proposed venue-informed degree measure. The results indicate that the venue-informed degree measure tended to be more efficient (smaller variance) and less biased than the other measure in both cities sampled. Venue attendance-adjusted network size may provide a more reliable and accurate degree measure for RDS estimates of the outcomes of interest.



Source: Fujimoto, K., Cao, M., Kuhns, L. M, Li, D. H., & Schneider, J. A. (2018). Statistical adjustment of network degree in respondent-driven sampling estimators: Venue attendance as a proxy for network size among young men who have sex with men. Social Networks, 54, 118–131.

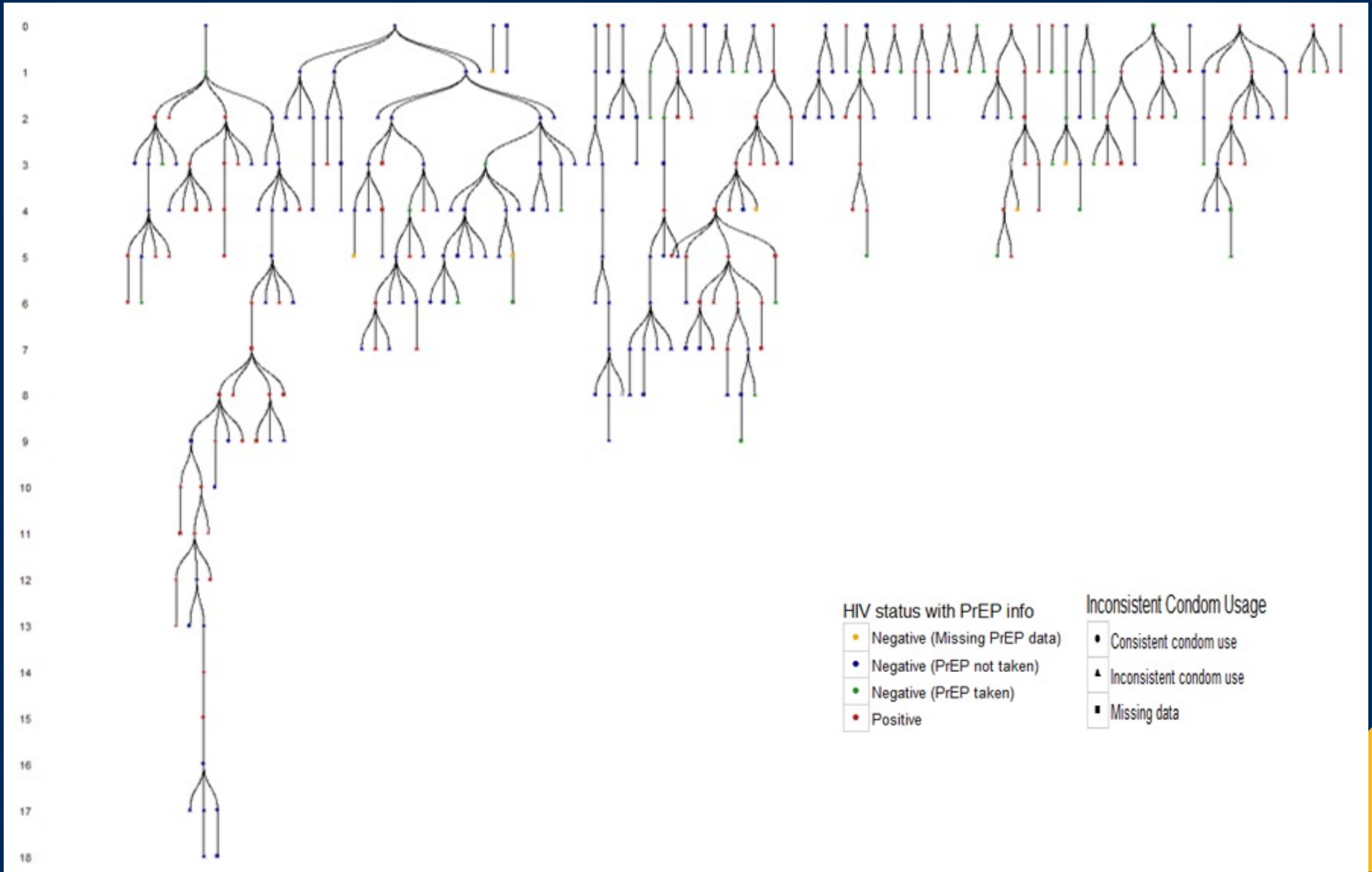
# RDS Chains among YMSM for Chicago, excluding non-sprouted seeds

(The longest recruitment chain per site was 9 waves, with 34 seeds (constituting 9.0% of the sample) recruited)



# **RDS Chains** among YMSM for **Houston**, excluding non-sprouted seeds

(The longest recruitment chain per site was 18 waves, with 63
seeds (constituting 16.7% of the sample) recruited)



# Statistical network model: Exponential Random Graph Modeling (ERGM)

- Statistical method of directly modeling observed network
- Test whether there are statistically significant structural characteristics in the observed network
  - Assess if these characteristics are more likely to be observed in the network than randomly generated networks

# Formulation: ERGMs

ERGMs take the form of probability distribution of graph (for all tie-variables simultaneously)

$$P(Y=y)=1/(k(\theta))\ \exp\{\theta\hat{}'\ g(y)\}$$

$Y$ is a set of tie indicator variables $Y$

$y$ is a realization, the observed network

$g(y)$ is a vector of network statistics

$\theta$ is a parameter vector corresponding to $g(y)$

$k(\theta)$ is a normalizing factor calculated by summing up

$\exp\ \{\theta\hat{}' g(y)\}$ over all possible network configurations

# Empirical application of GCN using YMAP data

*Journal of the American Medical Informatics Association*, 26(11), 2019, 1263–1271
doi: 10.1093/jamia/ocz070
Advance Access Publication Date: 14 June 2019
Research and Applications

AMIA

OXFORD

Research and Applications

## Network context matters: graph convolutional network model over social networks improves the detection of unknown HIV infections among young men who have sex with men

**Yang Xiang,[1] Kayo Fujimoto,[2] John Schneider,[3,4] Yuxi Jia,[1,5] Degui Zhi,[1] and Cui Tao[1]**

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA, [2]Department of Health Promotion & Behavioral Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA, [3]Departments of Medicine and Public Health Sciences, University of Chicago, Chicago, Illinois, USA, [4]Chicago Center for HIV Elimination, University of Chicago, Chicago, Illinois, USA, and [5]Department of Medical Informatics, School of Public Health, Jilin University, Jilin, China

Corresponding Author: Cui Tao, PhD, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Suite 600, 7000 Fannin Street, Houston, TX 77054, USA (Cui.Tao@uth.tmc.edu)

Received 3 October 2018; Revised 11 April 2019; Editorial Decision 25 April 2019; Accepted 29 April 2019

ABSTRACT

**Objective:** HIV infection risk can be estimated based on not only individual features but also social network information. However, there have been insufficient studies using n machine learning methods that can maximize the utility of such information. Leveraging a state-of-the-art network topology modeling method, graph convolutional networks (GCN), our main objective was to include network information for the task of detecting previously unknown HIV infections.

**Materials and Methods:** We used multiple social network data (peer referral, social, sex partners, and affiliation with social and health venues) that include 378 young men who had sex with men in Houston, TX, collected between 2014 and 2016. Due to the limited sample size, an ensemble approach was engaged by integrating GCN for modeling information flow and statistical machine learning methods, including random forest and logistic regression, to efficiently model sparse features in individual nodes.

**Results:** Modeling network information using GCN effectively increased the prediction of HIV status in the social network. The ensemble approach achieved 96.6% on accuracy and 94.6% on F1 measure, which outperformed the baseline methods (GCN, logistic regression, and random forest: 79.0%, 90.5%, 94.4% on accuracy, respectively; and 57.7%, 80.2%, 90.4% on F1). In the networks with missing HIV status, the ensemble also produced promising results.

**Conclusion:** Network context is a necessary component in modeling infectious disease transmissions such as HIV. GCN, when combined with traditional machine learning approaches, achieved promising performance in detecting previously unknown HIV infections, which may provide a useful tool for combatting the HIV epidemic.

Source: Xiang, Y., Fujimoto, K., Schneider, J. A., Jia, Y., Zhi, D., & Tao, C. (2019). Network context matters: Graph convolutional network model over social networks improves the detection of unknown HIV infections among young men who have sex with men. Journal of the American Medical Informatics Association (JAMIA), 26(11), 1263–1271.

# Overview of ensemble approach GCN for modeling information flow and statistical machine learning methods (random forest and logistic regression)

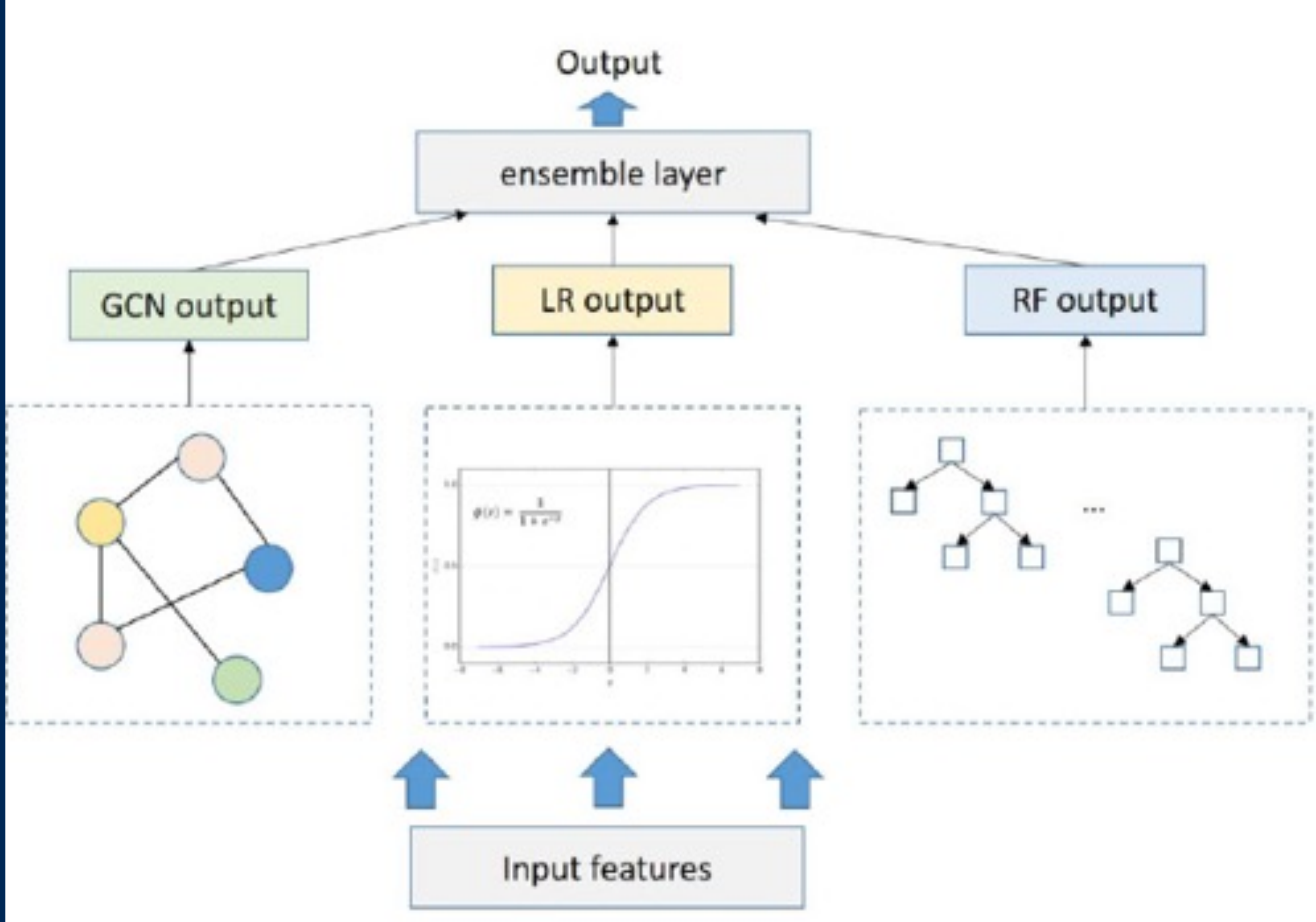



Table 2. Prediction performance for different classifiers in percent (standard deviation of 10 rounds of resampling)

| Method | Accuracy(%) | F1(%) |
|---|---|---|
| GCN | 79.0(±2.72) | 57.7(±6.83) |
| LR | 90.5(±2.69) | 80.2(±7.48) |
| RF | 94.4(±2.76) | 90.4(±4.38) |
| GCN+LR | 93.4(±3.09) | 88.4(±5.30) |
| GCN+RF | 96.6(±1.97) | 94.6(±2.88) |
| LR+RF | 95.3(±2.75) | 91.6(±5.64) |
| GCN+LR+RF | 96.5 (±2.05) | 94.5(±3.12) |

GCN effectively increased the prediction of HIV status in the social Network, achieved 96.6% on accuracy and 94.6% on F1 measure, which outperformed the baseline methods

# Empirical application of GAT using YMAP data

## Identifying influential neighbors in social networks and venue affiliations among young MSM: a data science approach to predict HIV infection

Yang Xiang[a], Kayo Fujimoto[b], Fang Li[a], Qing Wang[a], Natascha Del Vecchio[c], John Schneider[c,d], Degui Zhi[a] and Cui Tao[a]

**Objective:** Young MSM (YMSM) bear a disproportionate burden of HIV infection in the United States and their risks of acquiring HIV may be shaped by complex multilayer social networks. These networks are formed through not only direct contact with social/sex partners but also indirect anonymous contacts encountered when attending social venues. We introduced a new application of a state-of-the-art graph-based deep learning method to predict HIV infection that can identify influential neighbors within these multiple network contexts.

**Design and methods:** We used empirical network data among YMSM aged 16–29 years old collected from Houston and Chicago in the United States between 2014 and 2016. A computational framework GAT-HIV (Graph Attention Networks for HIV) was proposed to predict HIV infections by identifying influential neighbors within social networks. These networks were formed by multiple relations constituted of social/sex partners and shared venue attendances, and using individual-level variables. Further, GAT-HIV was extended to combine multiple social networks using multigraph GAT methods. A visualization tool was also developed to highlight influential network members for each individual within the multiple social networks.

**Results:** The multigraph GAT-HIV models obtained average AUC values of 0.776 and 0.824 for Chicago and Houston, respectively, performing better than empirical predictive models (e.g. AUCs of random forest: 0.758 and 0.798). GAT-HIV on single networks also delivered promising prediction performances.

**Conclusion:** The proposed methods provide a comprehensive and interpretable framework for graph-based modeling that may inform effective HIV prevention intervention strategies among populations most vulnerable to HIV.



*AIDS* 2021, **35** (Suppl 1):S65–S73

Source: Xiang, Y., Fujimoto, K., Li, F., Wang, Q., *Del Vecchio, N., Schneider, J. A., Zhi, D., & Tao, C. (2021). Identifying influential neighbors in social networks and venue affiliations among young MSM: A data science approach to predict HIV infection. AIDS, 35(Suppl 1), S65–S73.

# Empirical application of organizational network in PrEP care delivery system using YMAP

Networks, Knowledge Brokers, and the Public Policymaking Process pp 265–314 | Cite as

## Brokerage-Centrality Conjugates for Multi-Level Organizational Field Networks: Toward a Blockchain Implementation to Enhance Coordination of Healthcare Delivery

Kayo Fujimoto, Camden J. Hallmark, Rebecca L. Mauldin, Jacky Kuo, Connor Smith, Natascha Del Vecchio, Lisa M. Kuhns, John A. Schneider & Peng Wang

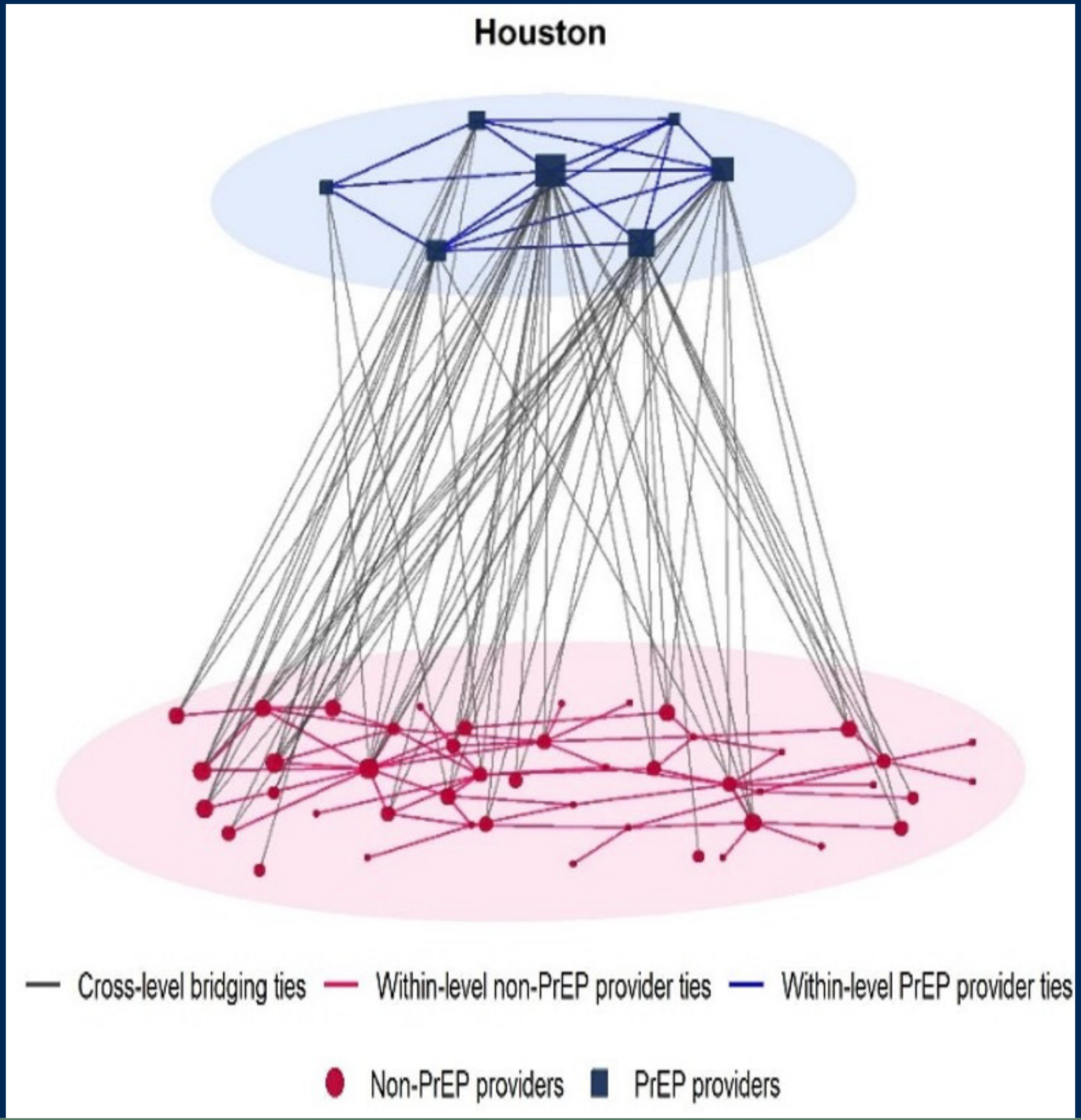
Chapter | First Online: 04 November 2021



### Abstract

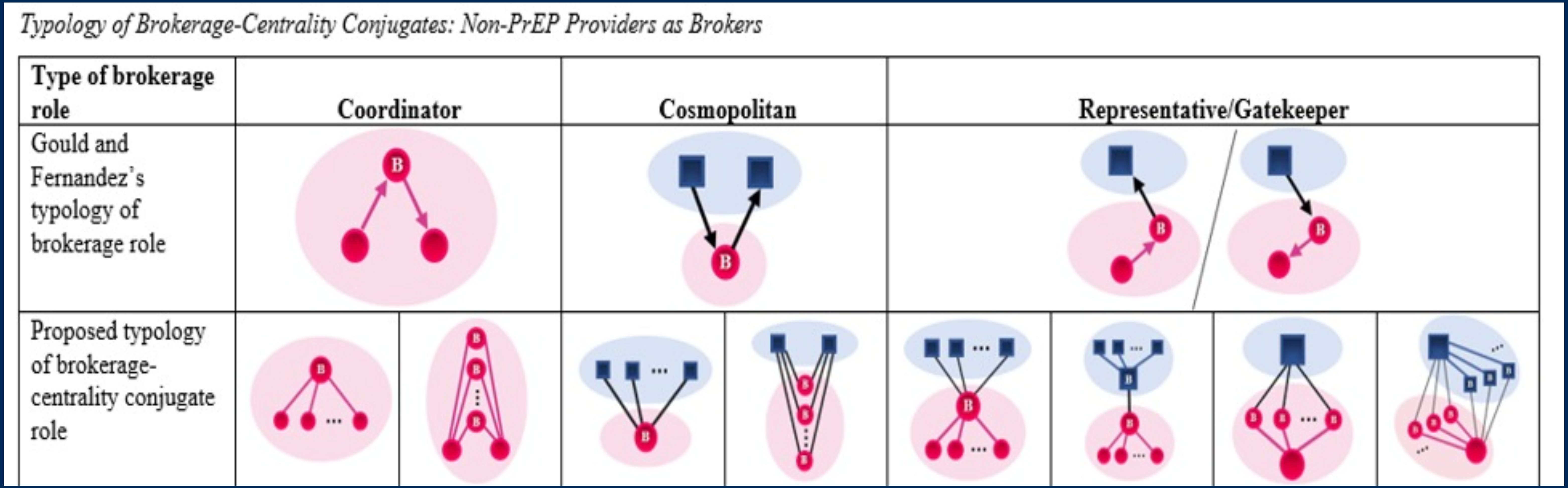
A fragmented U.S. healthcare delivery system may reflect a highly brokered communication network controlled by only a few brokers. Such relational inequality in brokerage influences the formation of interorganizational brokerage relations. This chapter presents the theoretical mechanisms that underlie the reasoning of instantiating organizational power dynamics in controlling communication (brokerage) while increasing connectivity (degree centrality); determining opportunities for accessing and exchanging resources across subgroups; and shaping organizational decisions and actions that are transformed collectively into locally centralized or decentralized brokerage structures, called 'brokerage-centrality conjugates.' These conjugates make up the interorganizational collaboration network of an organizational field, and is tested and supported by multi-level exponential random graph models. Finally, a blockchain-based network intervention is proposed to enhance interorganizational communication and coordination to improve population health.

# Multilevel collaboration network in PrEP care delivery system




Source: Fujimoto, K., Hallmark, C. J., Mauldin, R. L., Kuo, J. C., Smith, C., Del Vecchio, N., Kuhns, L. M., Schneider, J. A., & Wang, P. (In Press). Brokerage-centrality conjugates for multi-level organizational field networks: Toward a blockchain implementation to enhance coordination of healthcare delivery. (Eds.) Weber, M. S., & Yanovitzky, I. Networks, Knowledge Brokers, and the Public Policymaking. Springer International Publishing AG, Cham.

# Local structural configurations (motifs) Brokerage structures by Gould & Fernandez (1989) and our extension

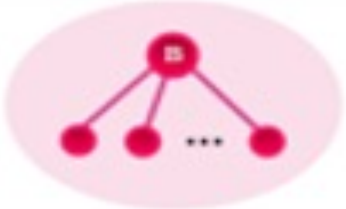
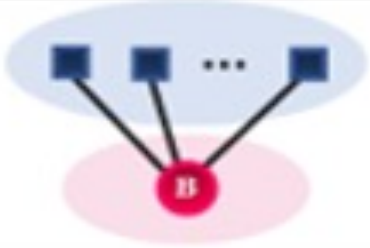
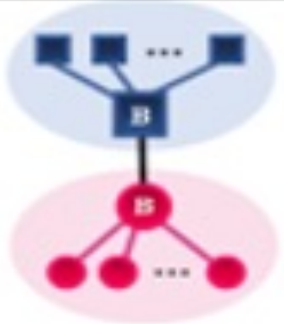



Letter B in the circles or squares represent broker organizations(s).

These were all significant positive estimates,  indicating that these structural configurations are more likely to be observed in the network than randomly generated networks.

# Results of important structural configurations that drive collaboration network in PrEP delivery system

We observe some bottleneck structures with a single broker or multiple brokers connected through the same pair of organizations.

| Configurations | Houston | |
|---|---|---|
| | Parameter | SE |
|  | 1.084 | 0.373 |
|  | 0.096 | 0.015 |
|  | 1.492 | 0.411 |
|  | 0.121 | 0.012 |
|  | 0.095 | 0.014 |

# Implementation challenges in local PrEP care delivery system in Houston

- At the individual level: PrEP uptake among most-at-risk younger MSM of color is low
    - Social determinants of health such as unemployment, housing instability, incarceration, stigma, discrimination.

medical mistrust and distrust.

- At the organizational level: The existing client referral system generate bottleneck and concentration of information and resources in the local PrEP referral network
    - PrEP referrals are made based on existing relationships with specific providers (i.e., within-organizational referral), or a few prominent providers or health agencies

      → generate hubs of information and resource flows, being incapable of handling sudden environmental changes (e.g., the COVID- 19 pandemic)
- At the community level: The existing client referral system may drive fragmented local PrEP care delivery system.

# Technological solution for PrEP implementation challenge: “Blockchain-based network intervention”

## What is Blockchain?

- Distributed ledger technology,
- Highly secured, decentralized database system that is maintained by every participant in a peer-to-peer (P2P) network without central authority as an intermediary.
- Decentralized governance (consensus algorithms) and shared ownership.

UTHealth Houston
School of Public Health

# Blockchain Project Aims/Objectives

- Aim 1:
  - To develop a schema to capture the data elements that will be required to report HIV status information
  - Time, date, type, location, result, of the HIV test).
- Aim 2:
  - To identify antecedent and anticipated implementation outcomes:
    - Acceptability, feasibility, adoptability, implementability
    - Potential determinants of implementing TestLinker among community stakeholders

# Acknowledgement

This research was funded by:

National Institutes of Health (P30AI161943; 1R01MH100021)

Houston Health Department

The Sally W. Vernon, Ph.D. Distinguished Professorship in Social Determinants of Health, UTHealth Houston

This research was consulted by:

R3EDI Implementation Science hub, Yale CIRA

# Thanks for your attention
# Any ideas?

Kayo Fujimoto, PhD, MS

Kayo.Fujimoto@uth.tmc.edu

Texas Developmental Center for AIDS Research
TX D-CFAR

UTHealth® Houston
School of Public Health

HEALTH
WITHOUT
BOUNDARIES

sph.uth.edu