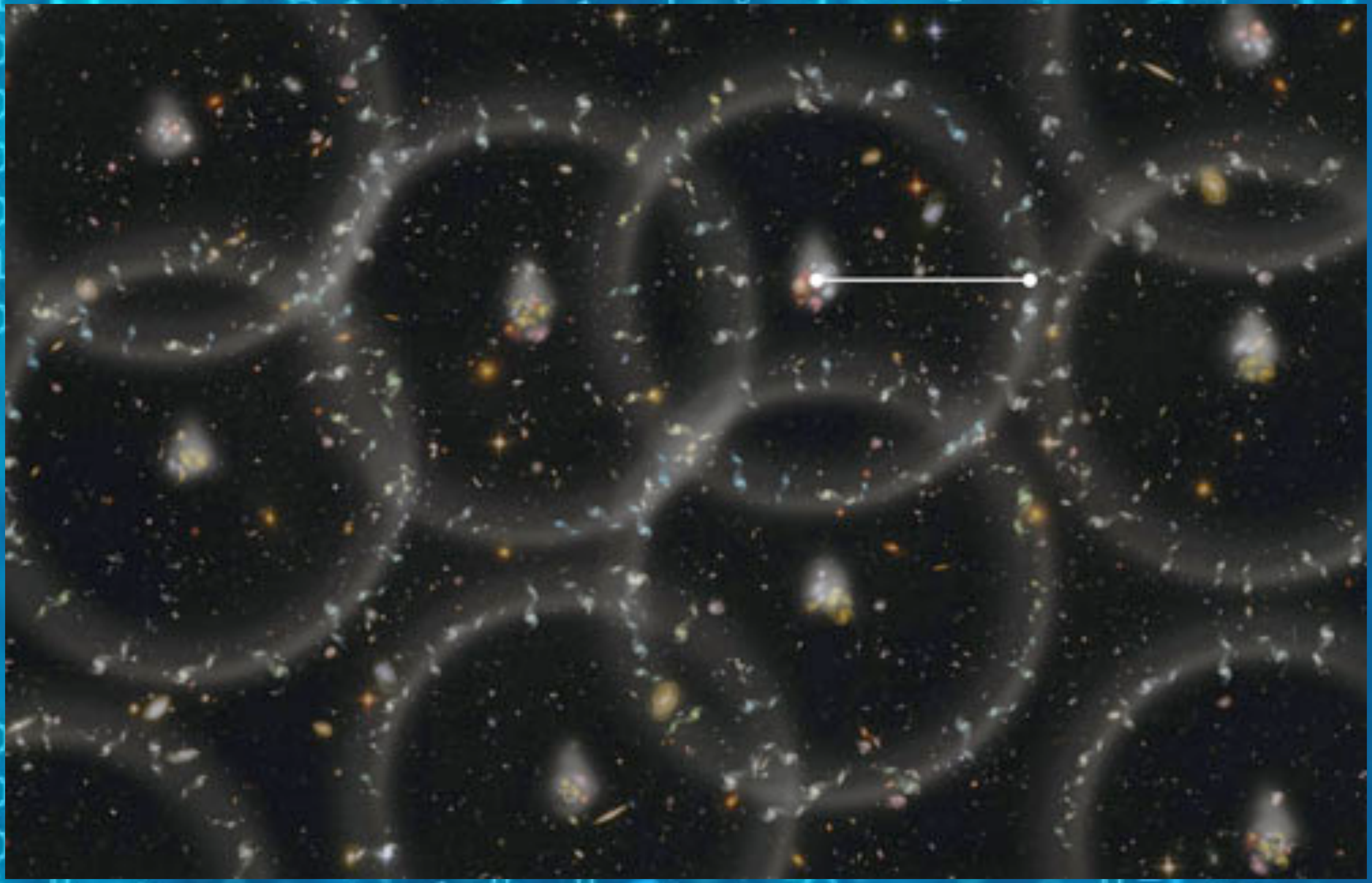
# Graph Databases Solution for Higher Order Statistics in the Big Data Era of Astronomy

**Cristiano Sabiu（Yonsei University）**

**Juhan Kim（KIAS）, Xiao-Dong Li（Sun Yat-Sen）, Ben Hoyle（LMU）**



**CosKASI conference III “The Correlated Universe”, Jeju, April 23-27**

# Contents

- Graph Databases
    - Introduction
    - Higher Order Spatial Clustering
    - Constructing Graph
    - Querying Graph

- Applications
    - Baryon Acoustic Oscillations
    - Cross-correlation with galaxy properties
    - Early Universe and Isocurvature Perturbations

- Conclusions

# Graph Databases

A graph within graph databases is based on graph theory.

- Nodes represent entities or instances such as people, businesses, accounts, or any other item to be tracked. They are roughly the equivalent of the record, relation or row in a relational database, or MySQL item

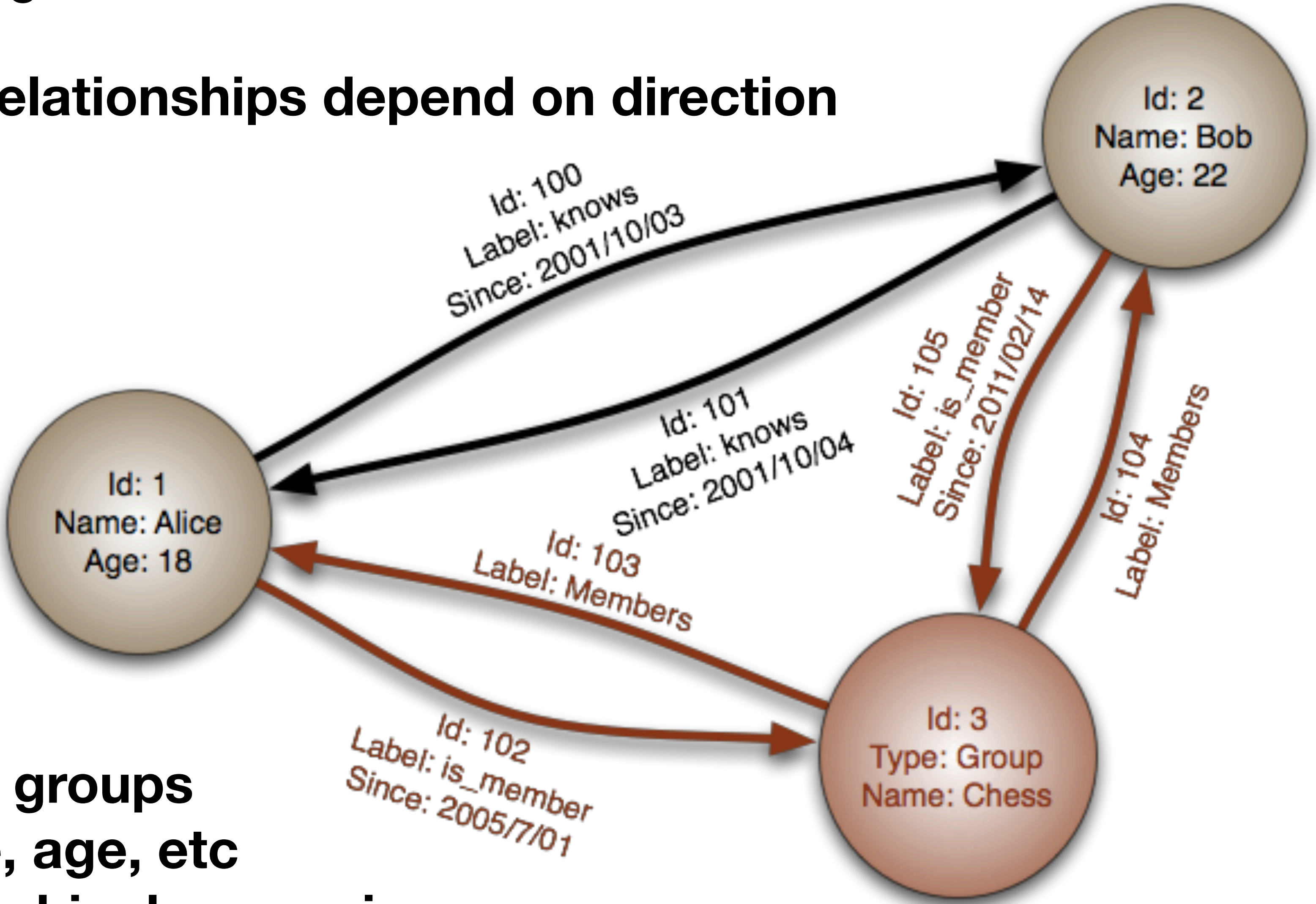
- Relationships, are the lines that connect nodes to other nodes; representing the relationship between them. Meaningful patterns emerge when examining the connections and interconnections of nodes, properties and relationships. Relationships are the key concept in graph databases, representing an abstraction that is not directly implemented in a relational model like SQL

- Properties are germane information to nodes. For example, if *Wikipedia* were one of the nodes, it might be tied to properties such as website, reference material, or words that starts with the letter *w*, depending on which aspects of Wikipedia are germane to a given database.

# Graph Databases

**In 2009 Facebook gave up using their MySQL storage and moved to a graph structure**

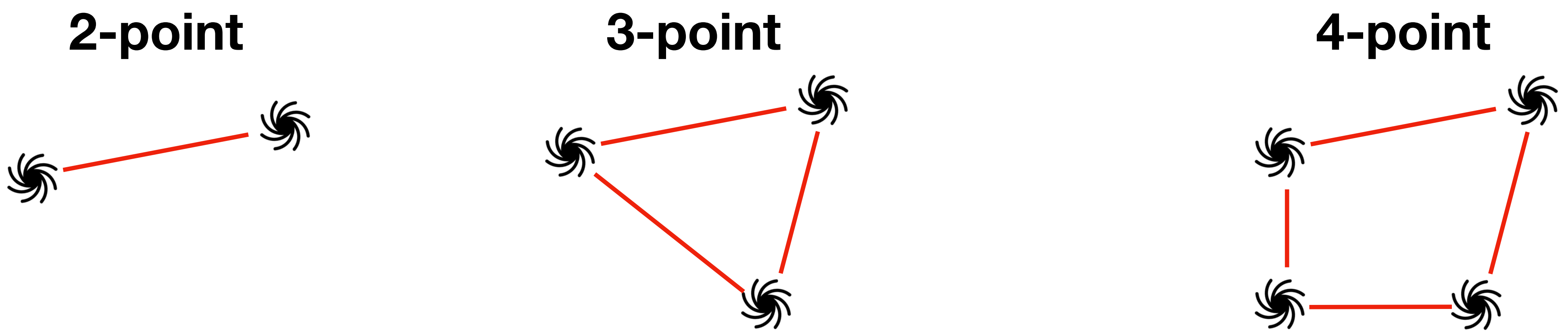
**A directed graph: relationships depend on direction**



**Nodes: people, places, groups**
**Node properties: name, age, etc**
**Relationships: membership, known since**

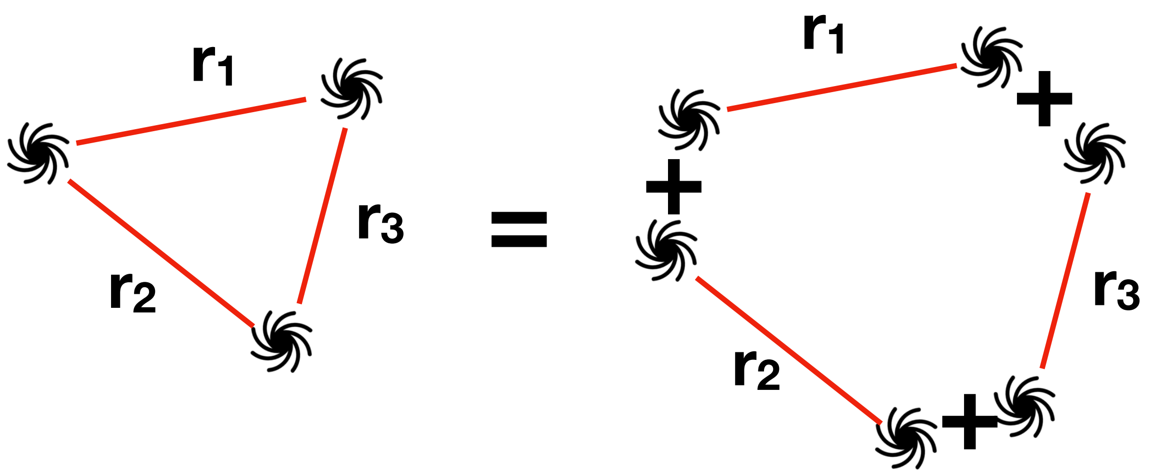
**Graph for our purposes will be much simpler!**

# Graph Database solution for Galaxy Clustering statistics



- The computational expensive part of measuring higher order statistics is in making sure the distances between data points satisfy our specific criteria i.e. $r_1$, $r_2$, $r_3$, etc

- However we notice that all higher order statistics are just complex combinations 2-point statistics



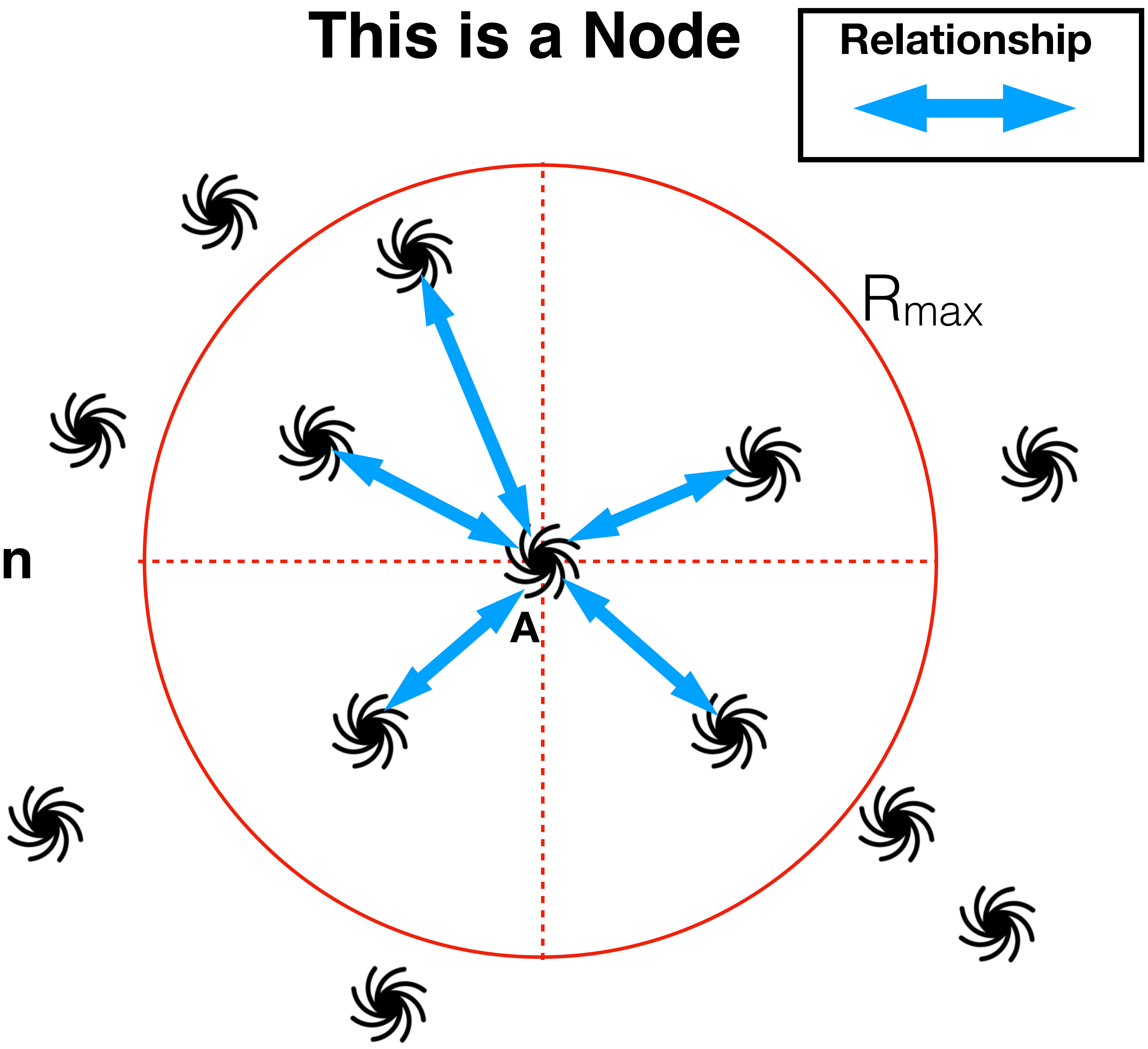
So rather than treat data positions as the important feature, rather treat each data pair (relationship) as the main information!

# Graph Database solution for Galaxy Clustering statistics

Each galaxy (or random) point is a node which may have relationships to there nodes

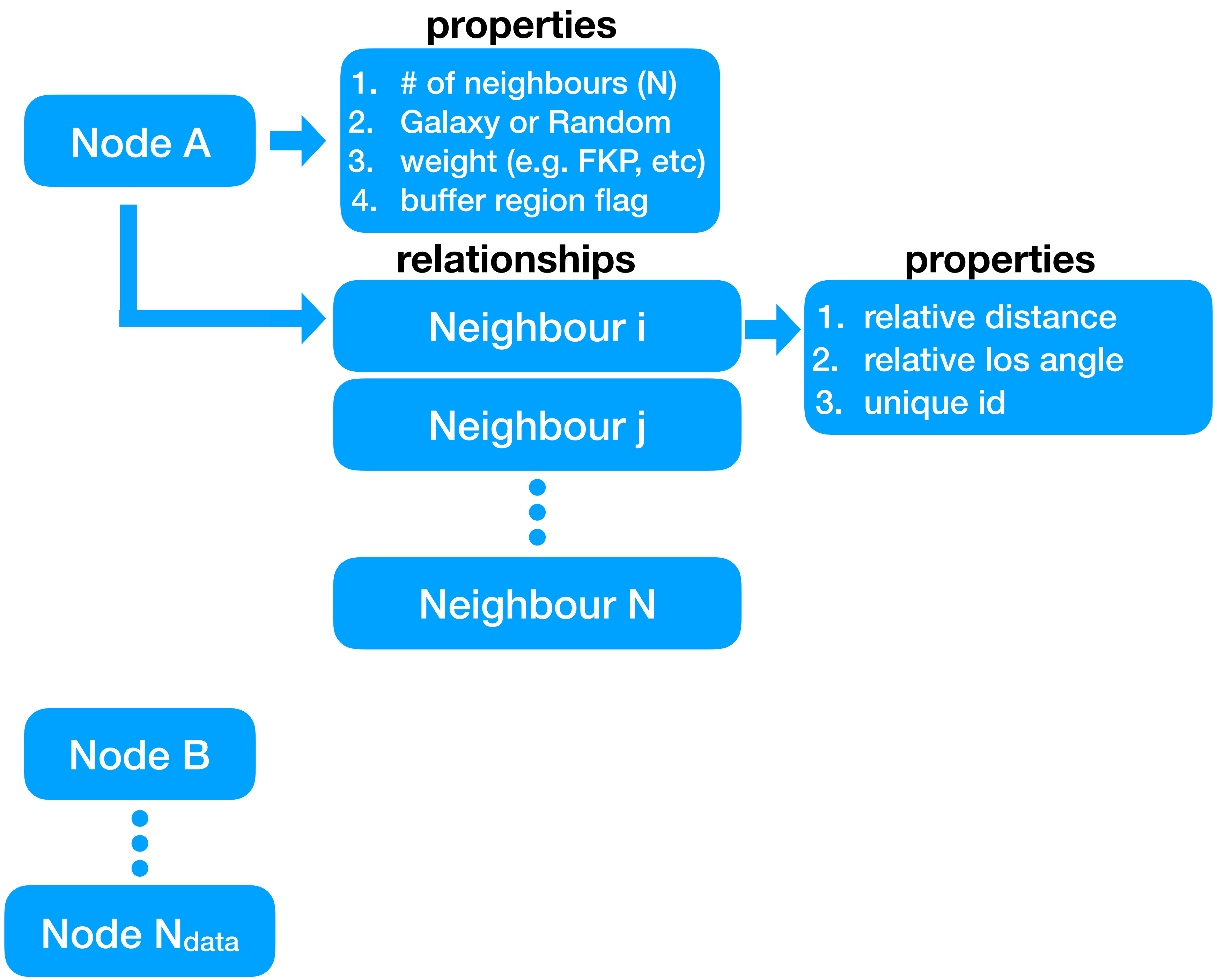
For our purposes the relationship information is the distance to neighbours within $R_{max}$

Neighbour list can be obtained quickly using a kd-tree algorithm



# Graph Database solution for Galaxy Clustering statistics

## Graph Database Structure



Node A

properties

1. # of neighbours (N)
2. Galaxy or Random
3. weight (e.g. FKP, etc)
4. buffer region flag

relationships

Neighbour i

Neighbour j

Neighbour N

properties

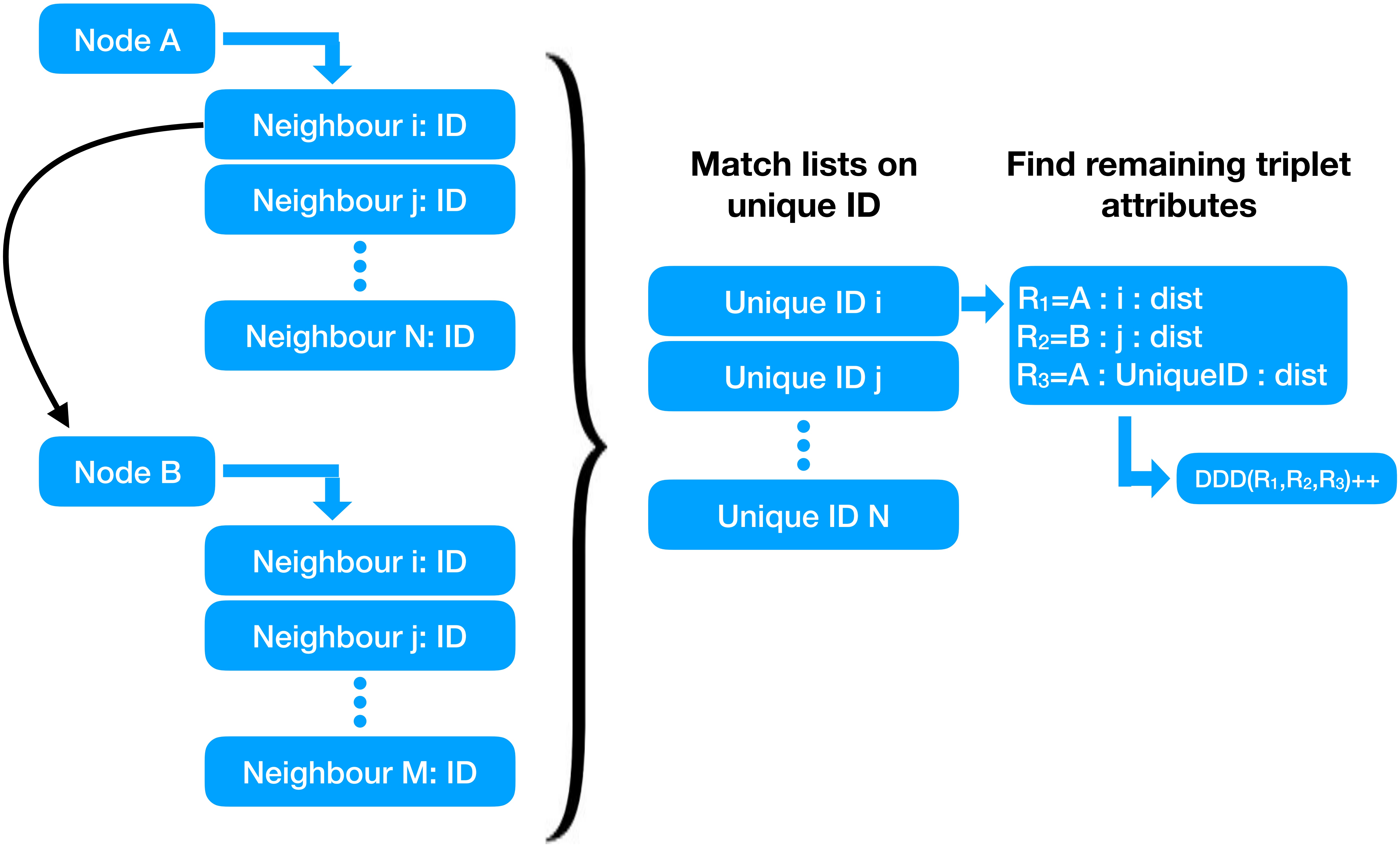
1. relative distance
2. relative los angle
3. unique id

Node B

Node $N_{data}$

# Graph Database solution for Galaxy Clustering statistics

## Query Graph: 3PCF



Node A

Neighbour i: ID

Neighbour j: ID

Neighbour N: ID

Node B

Neighbour i: ID

Neighbour j: ID

Neighbour M: ID

Match lists on unique ID

Unique ID i

Unique ID j

Unique ID N

Find remaining triplet attributes

$R_1$=A : i : dist
$R_2$=B : j : dist
$R_3$=A : UniqueID : dist

DDD($R_1$,$R_2$,$R_3$)++

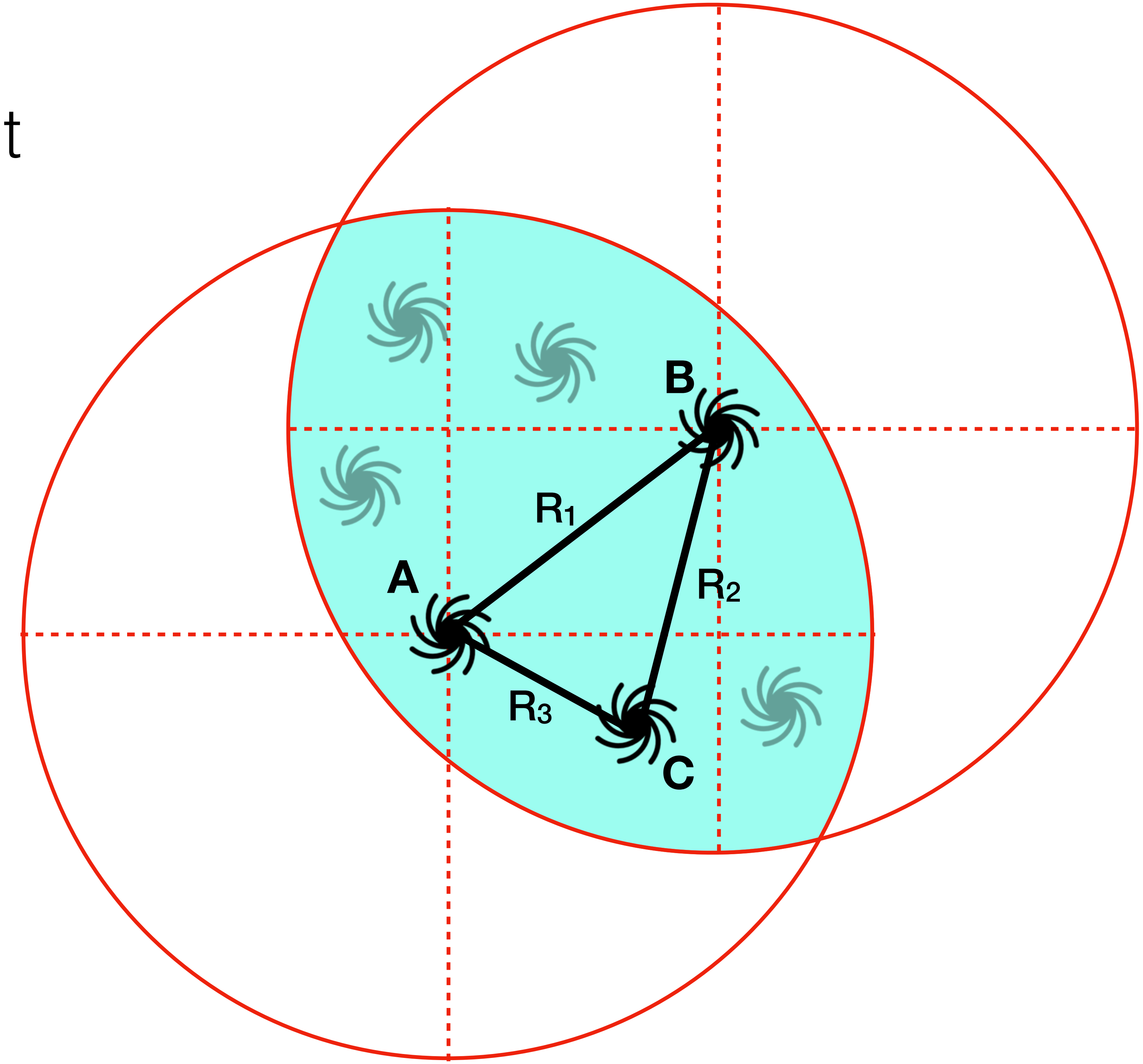
# Graph Database solution for Galaxy Clustering statistics

## Query Graph: 3PCF

The intersection of two sets (**A**,**B**) contains a list of points, **C**, that will complete the triangle A,B,C such that, **$0<R_1,R_2,R_3<R_{max}$**

The intersection must be computed ~ $N^2$ times!

For 2 **ordered lists**, the intersection can be computed very quickly!



# Graph Database solution for Galaxy Clustering statistics

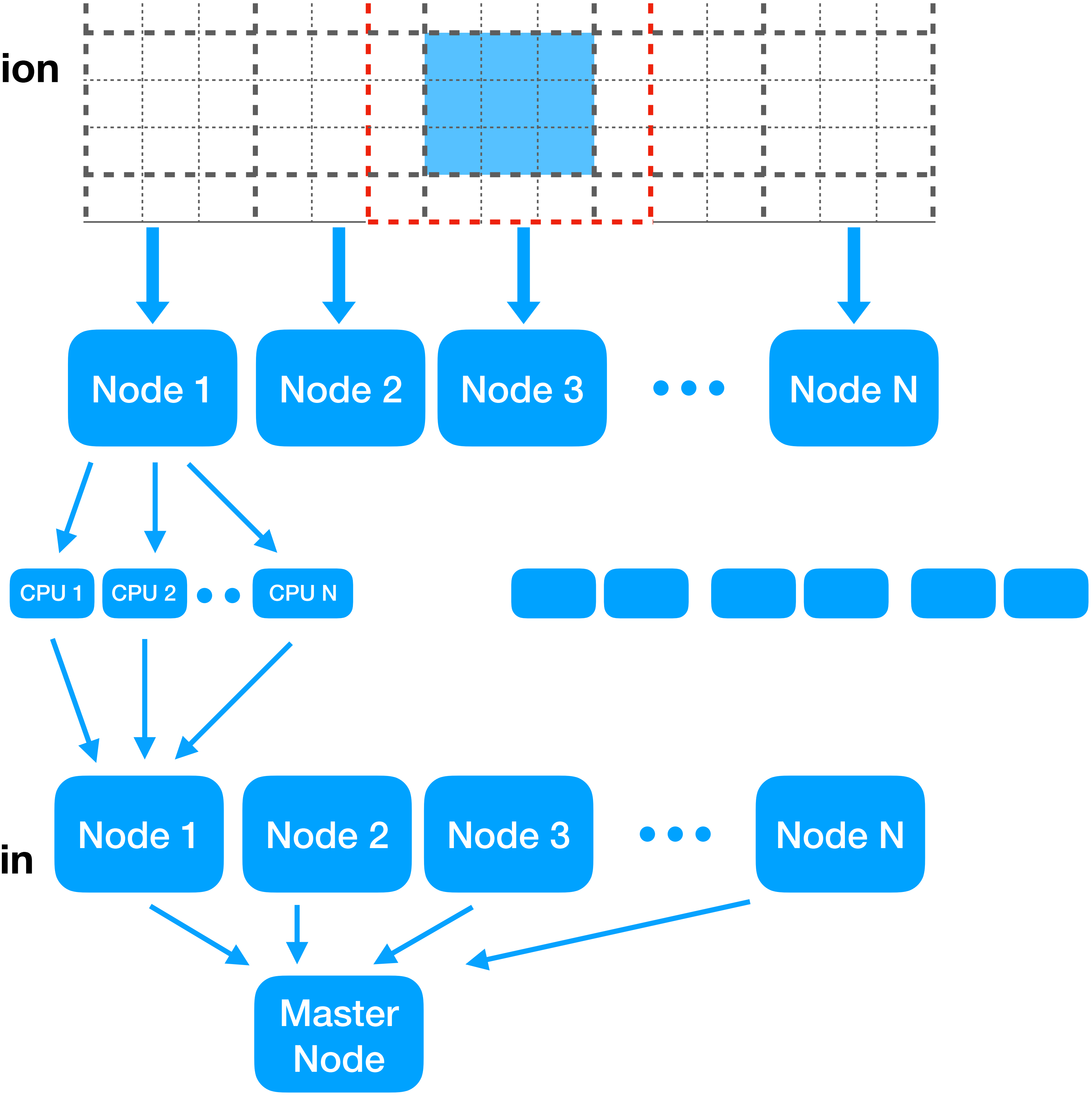
Domain decomposition is precomputed

Spawn *N* MPI processes

OpenMP threading

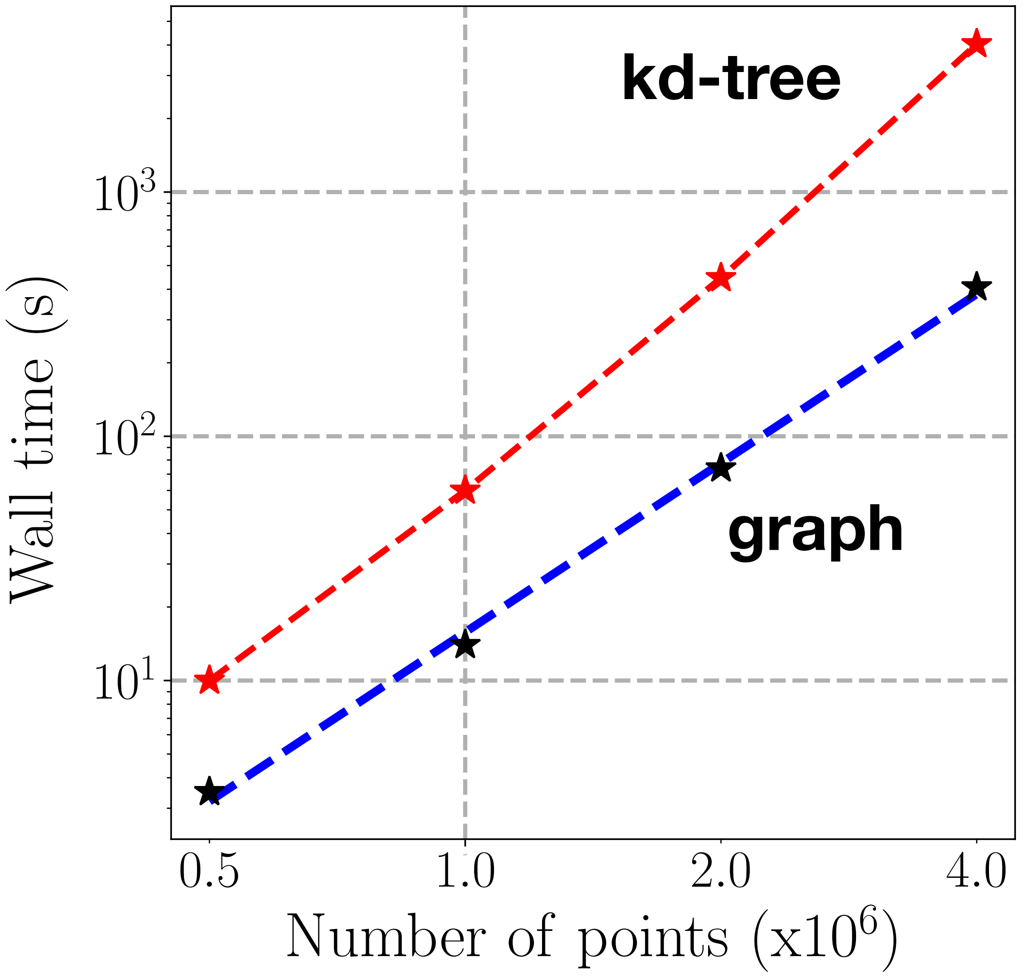
Reduce OMP arrays for each node/domain

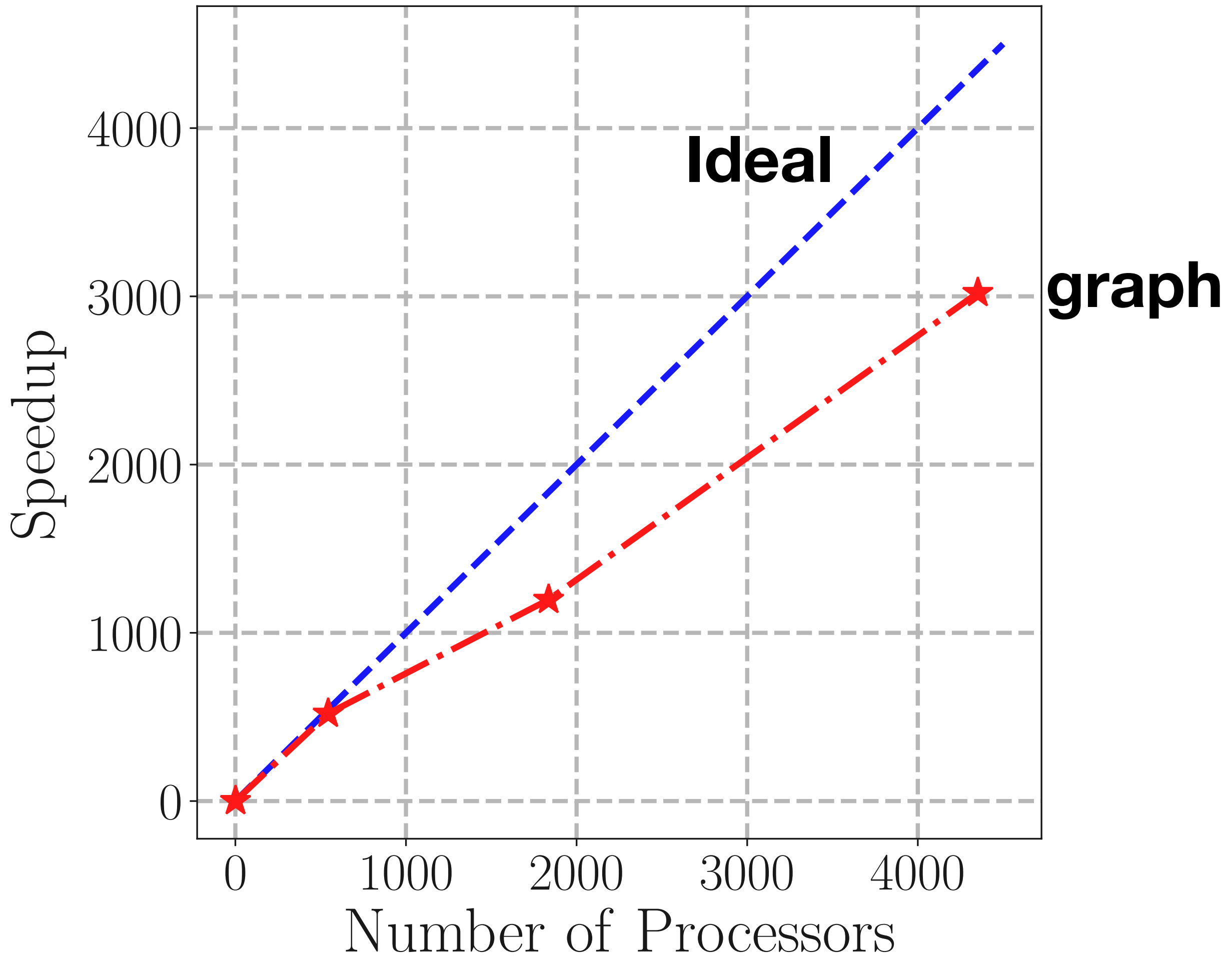
Collect MPI arrays and output



# Graph Database solution for Galaxy Clustering statistics

## Benchmarking





★C. Sabiu, B. Hoyle, J. Kim, X-D Li

★https://arxiv.org/abs/1901.00296

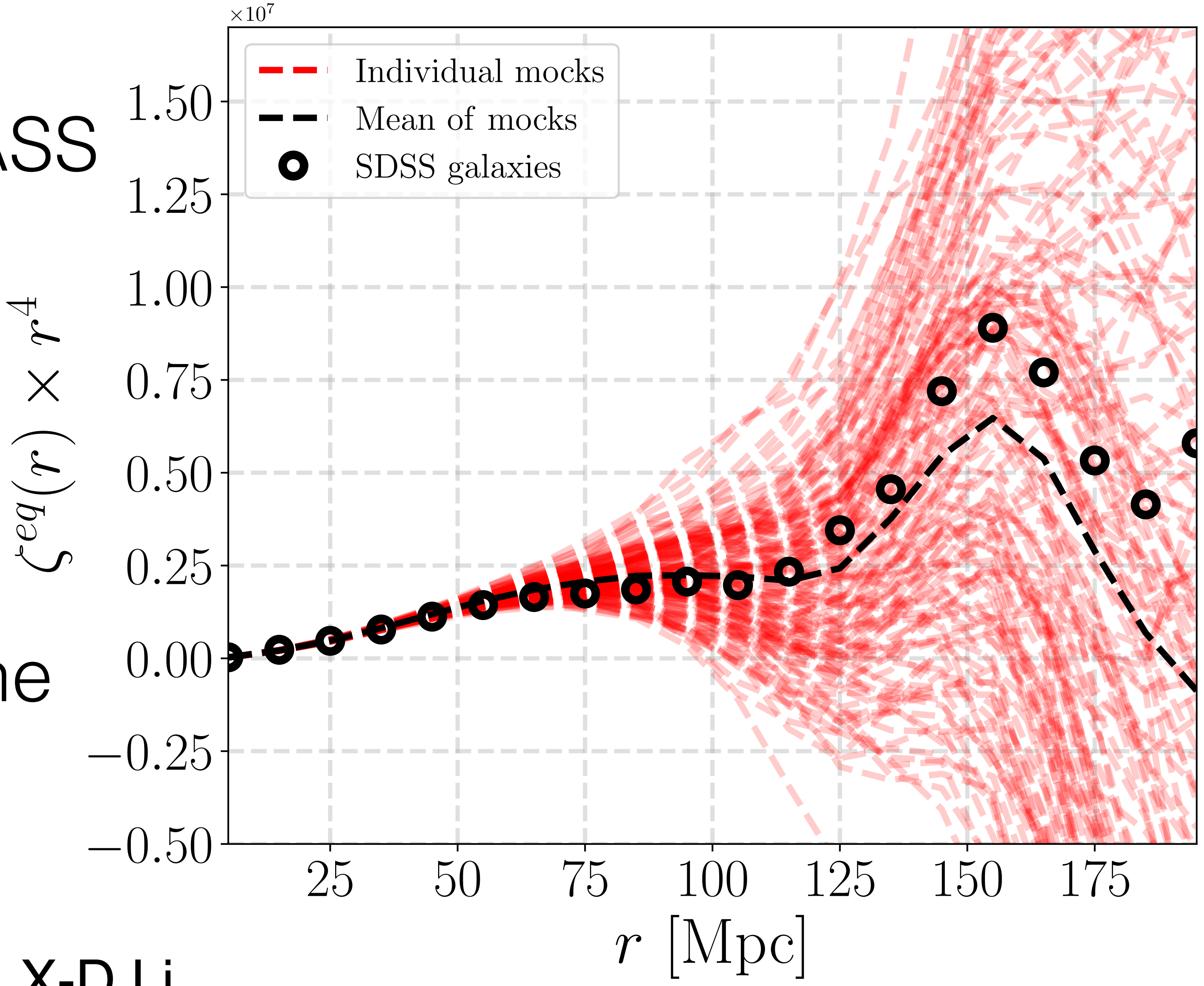
# Graph Database solution for Galaxy Clustering statistics

## BAO in the 3-point function

Taking the SDSS CMASS DR12 galaxies (~1M)

We look at equilateral triangles

We see evidence of the BAO peak



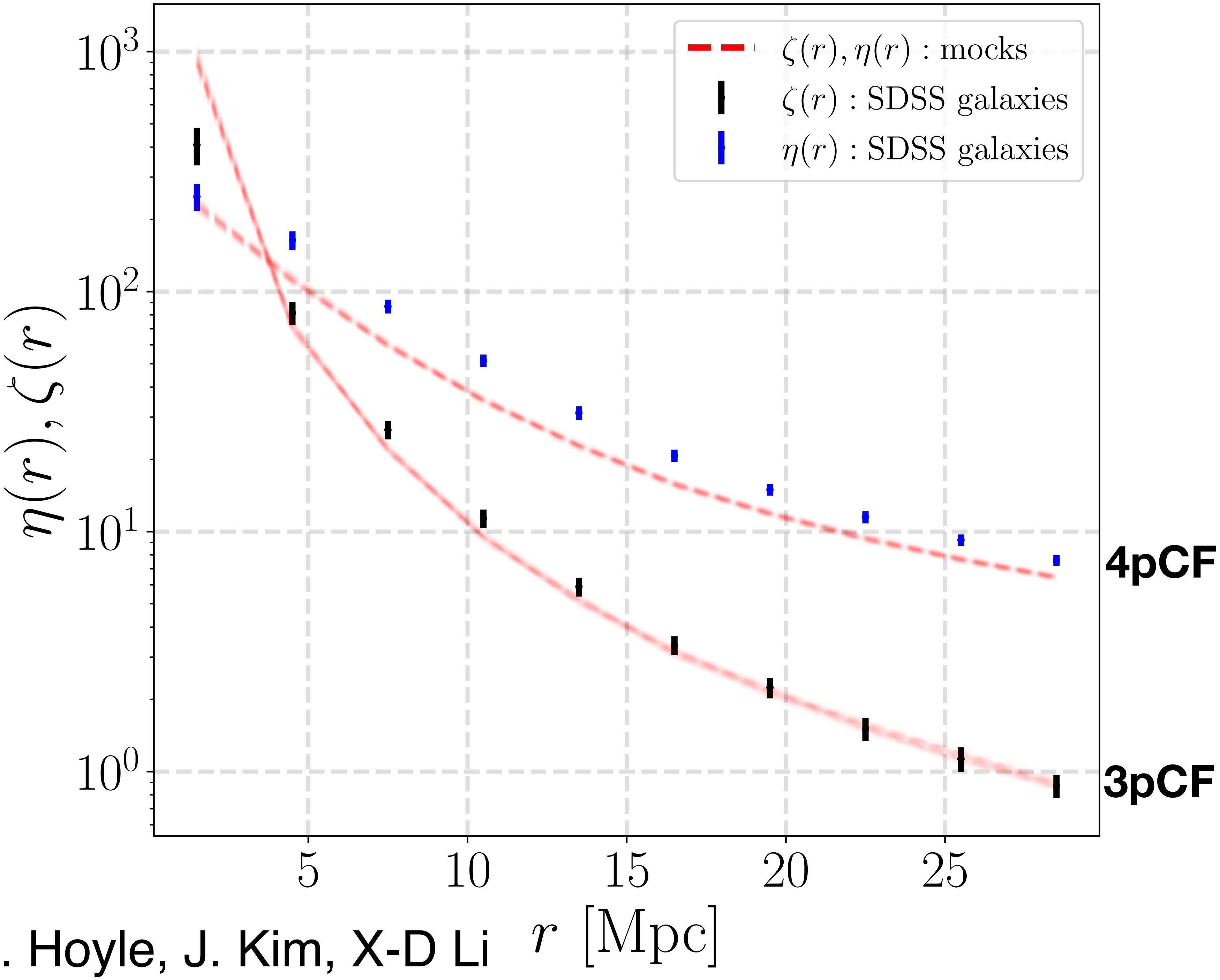
Isotropic, equilateral 3PCF

★C. Sabiu, B. Hoyle, J. Kim, X-D Li

★https://arxiv.org/abs/1901.00296

# Graph Database solution for Galaxy Clustering statistics

## 4-point correlation function

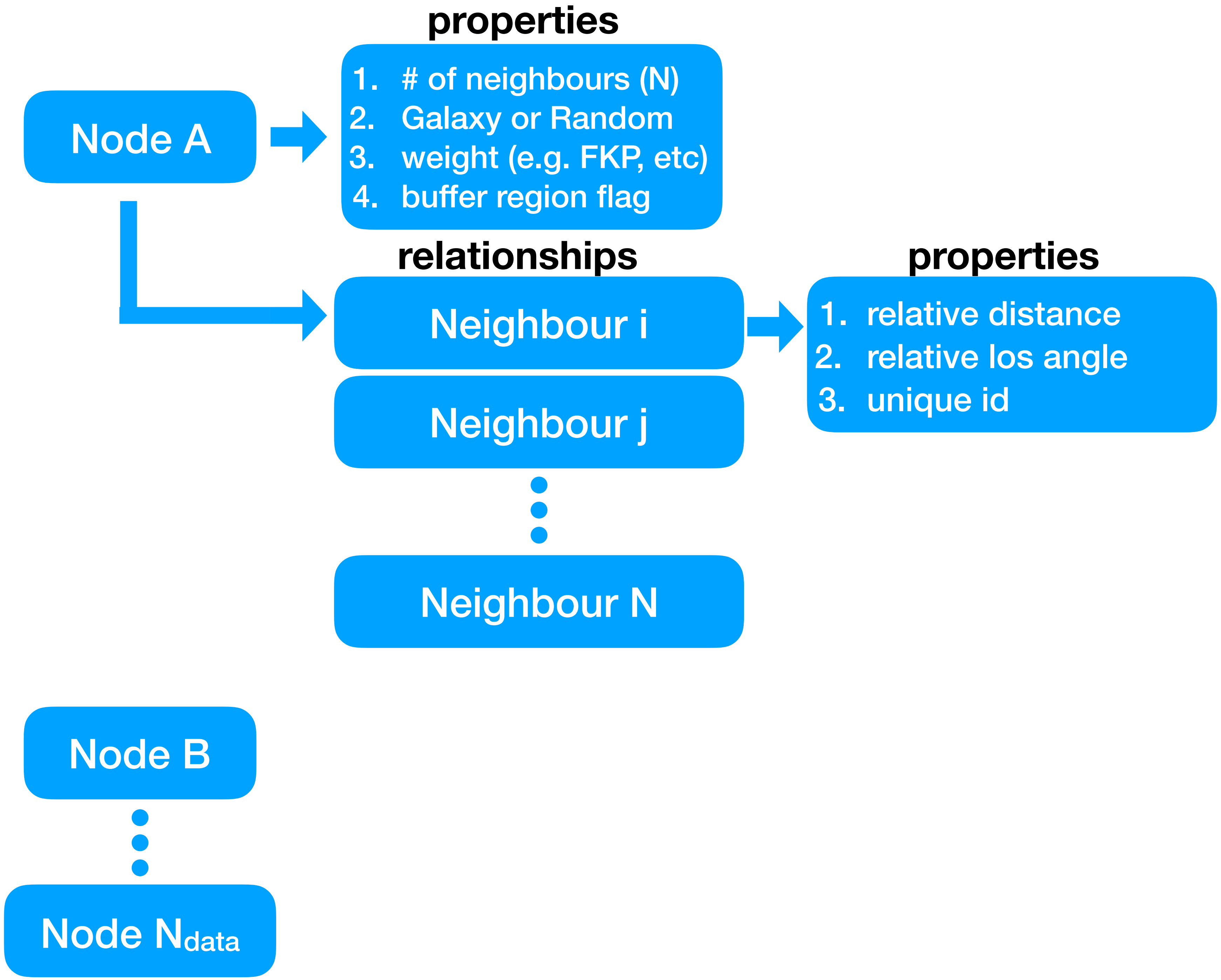


★C. Sabiu, B. Hoyle, J. Kim, X-D Li

★https://arxiv.org/abs/1901.00296

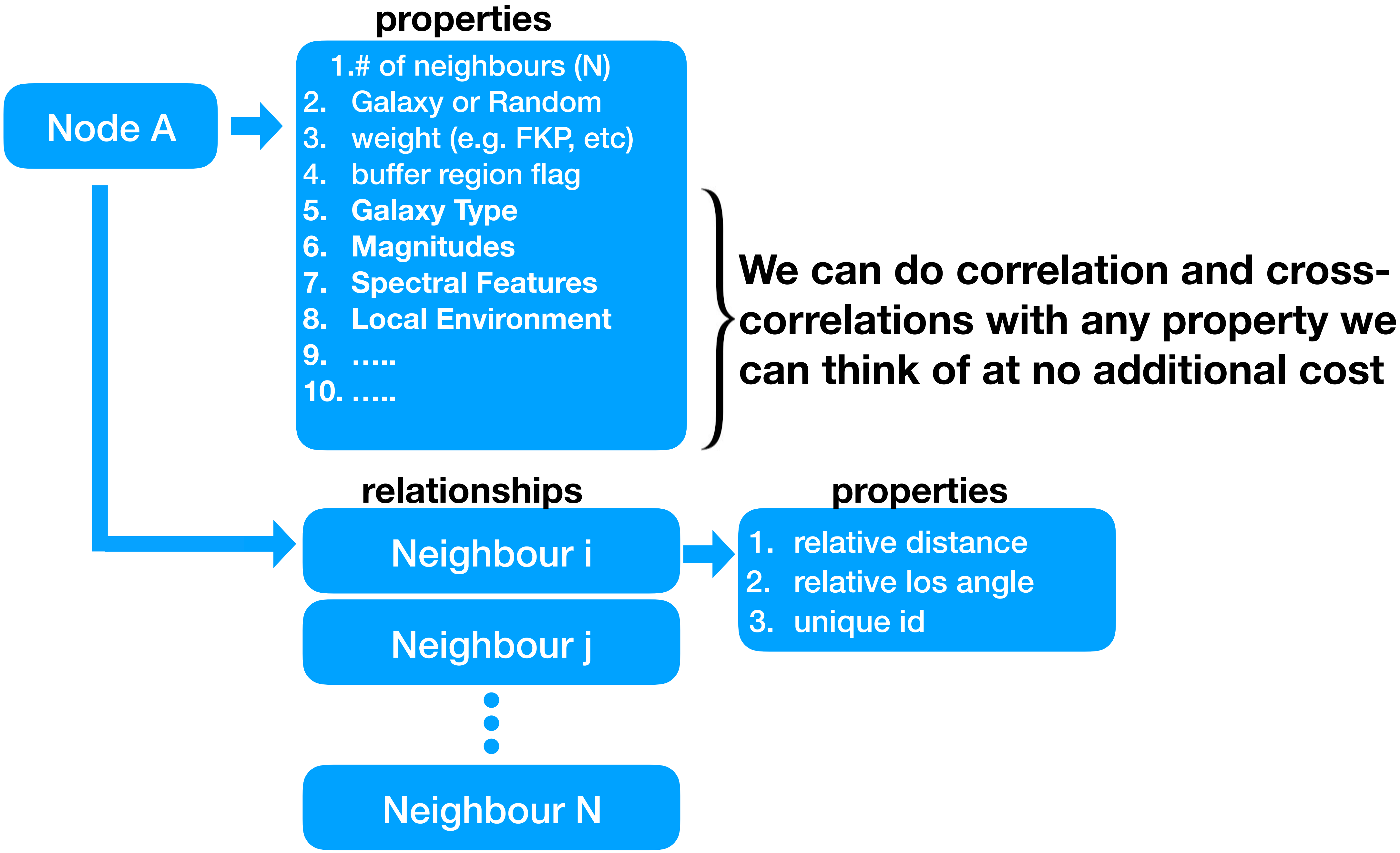
# Graph Database for exploring new quantities in galaxy evolution, cosmology

## Graph Database Structure



# Graph Database for exploring new quantities in galaxy evolution, cosmology

## Graph Database Structure

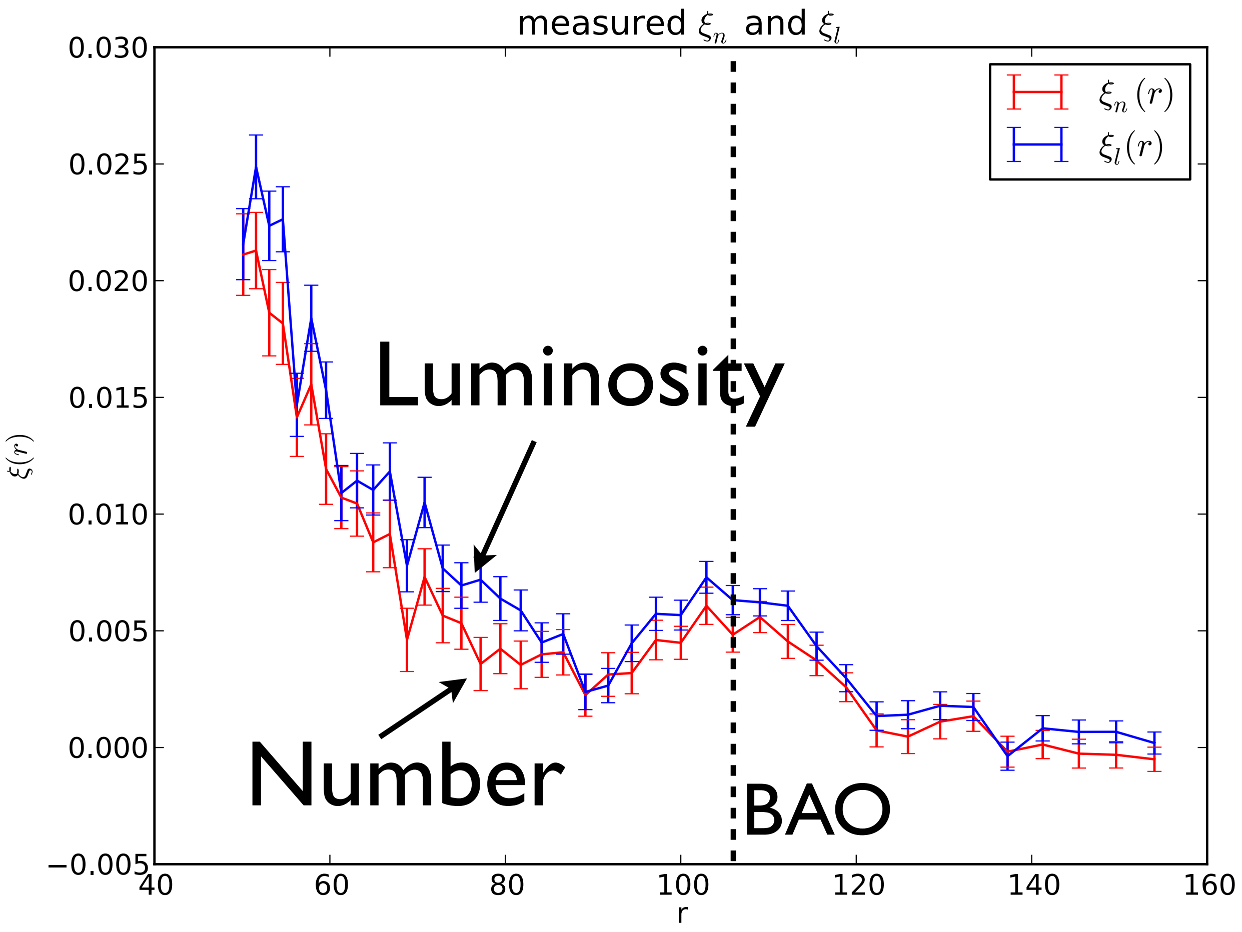


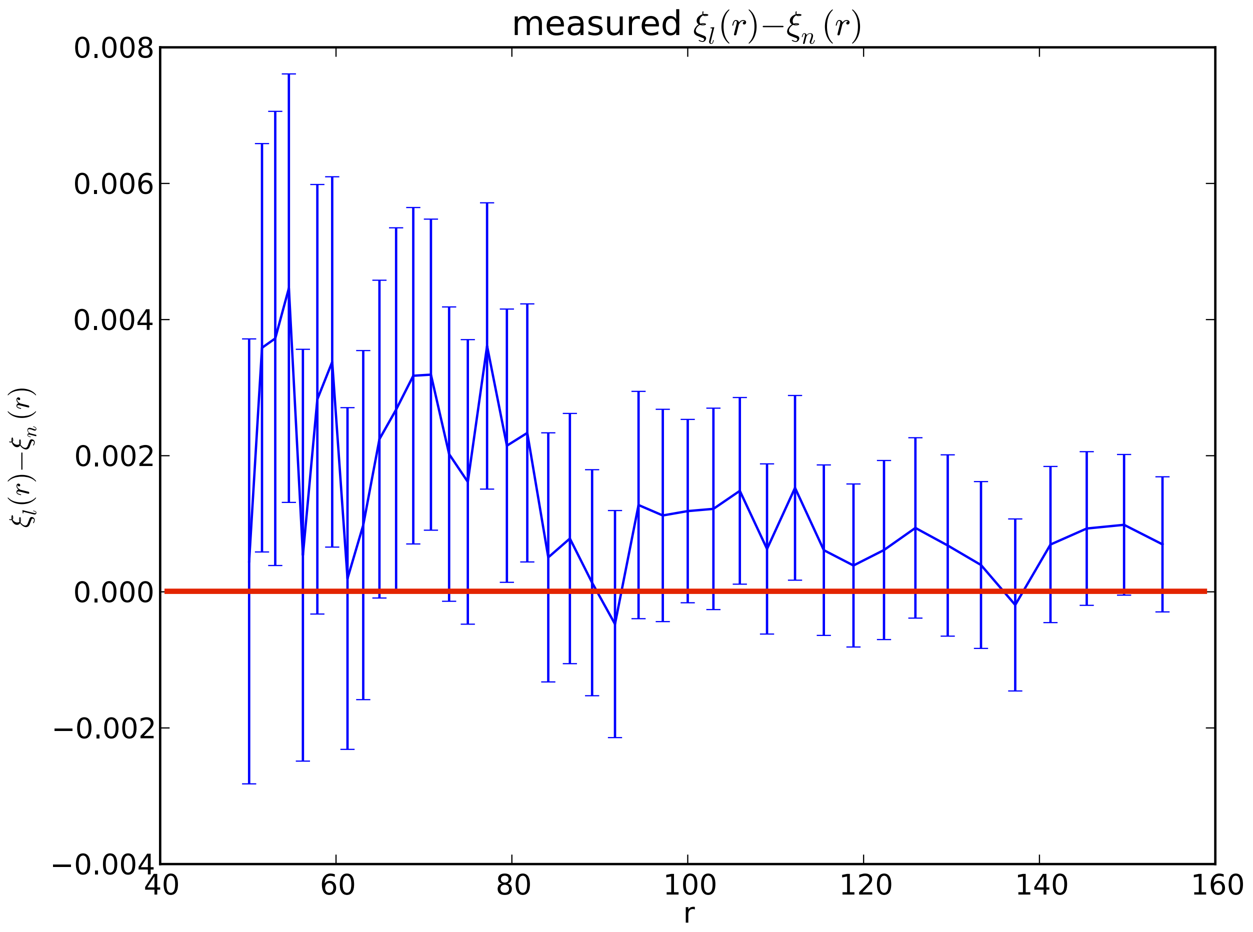
We can do correlation and cross-correlations with any property we can think of at no additional cost

# Graph Database for exploring new quantities in galaxy evolution, cosmology

## Example: lets look at counting data pairs while weighting by galaxy luminosity

- Query the database for all relationships where $r_{min}<r<r_{max}$
- Count that pair weighted by its absolute luminosity
- Compare with unclustered random points
- Compute the usual Landy-Szalay estimator for the 2pCF
- But what does it mean? What information does it contain?





# Graph Database for exploring new quantities in galaxy evolution, cosmology

**Following the theoretical work of Barkana & Loeb 2010**

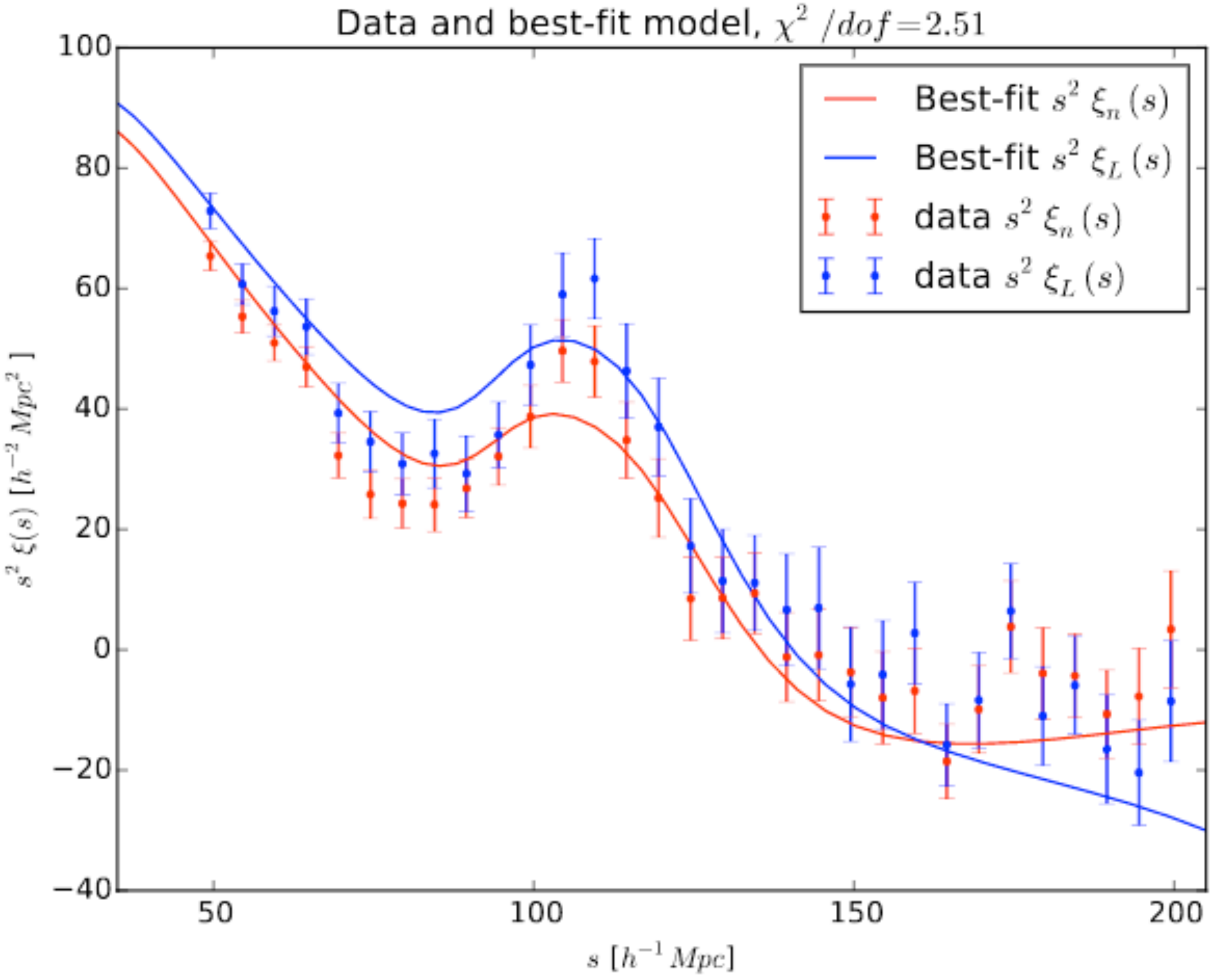
**We develop a model for the luminosity weighted correlation function of galaxies:**

$$\xi_L = B^2_{L,t}\xi_{\rm tot} + 2B_{L,t}B_{L,\Delta}\xi_{\rm add} + B^2_{L,\Delta}B_{\rm CIP}\hat{\xi}_{\rm CIP},$$

**This equation has dependence on:**

- **A linear bias with dark matter**
- **large scale clustering of baryons, potentially a new quantity to consider in galaxy evolution**
- **Compensated Isocurvature Perturbations (CIP) between baryons and dark matter in the early universe**

**Looking at spatial cross-correlations with different quantities unlocks new physical interpretation of the data**



M. T. Soumagnac, R. Barkana, C. G. Sabiu, A. Loeb, et.al **PRL 2016**
M. T. Soumagnac, C. G. Sabiu, R. Barkana, and J. Yoo **MNRAS 2019**

# Conclusions

GRAMSCI

v1 out soon!

- **We introduce a new clustering algorithm, soon publicly available under a GNU public release licence**

- **GRA**ph **M**ade **S**tatistics for **C**osmological **I**nformation: **GRAMSCI available soon from: http://bitbucket.org/csabiu/gramsci**

- **GRAMSCI performs much better than purely tree based approaches**

- **We show the performance by measuring all possible 3pCF unto and beyond the BAO scale with current SDSS BOSS data**

- **We make the first measurements of the 4-point function of SDSS galaxies**

- **We show the flexibility of adopting a Big Data Analytic approach. As an example the luminosity-number density cross correlation has the potential to unlock new information in the galaxy data that depends on baryonic physics and compensated isocurvature originating in the very early Universe.**

**Thank You 감사합니다**