Decentralized decision making on networks with FDR control

AADITYA RAMDAS, JIANBO CHEN, MARTIN J. WAINWRIGHT, MICHAEL I. JORDAN

Departments of Statistics and EECS, University of California, Berkeley

PROBLEM

- We consider a novel setting where distinct agents reside on the nodes of an undirected graph, and each agent possesses p-values corresponding to one or more hypotheses local to its node.
- Each agent must individually decide whether to reject one or more local hypotheses by only communicating with its neighbors.
- The goal is to control the global FDR over the entire graph at a predefined level.
- An example: Each node of the graph represents a sensor over a widespread area, such as a forest. Each sensor collects its own local data, but due to power constraints, each sensor may only communicate locally with nearby sensors. A rejected hypothesis reflects a discovery, like a fire or pollution-spike.

NOTATIONS

- $G = (\mathcal{V}, \mathbb{E})$, a graph with an agent at each node who may communicate with neighbors.
- $\mathcal{H}_i = \{H_{i,1}, \ldots, H_{i,n_i}\}$: the set of n_i hypotheses being tested by the agent at node i.
- $\mathcal{H}_i^0$ represent the (unknown) subset of true null hypotheses at node i.
- $\mathcal{H}^0$ represent the overall set of true null hypotheses.
- P-values at each nodes: $P_{i,1}, \ldots, P_{i,n_i}$.
- R, the total number of discoveries.
- V, the total number of false discoveries.

PRELIMINARIES

- False Discovery Rate (FDR): $\text{FDR} = \mathbb{E}\left[\frac{V}{R}\right]$.
- Positive Regression Dependence on a Subset (PRDS): For any $i \in \mathcal{H}^0$ and nondecreasing set $\mathcal{D} \subseteq [0,1]^n$, the function $t \mapsto \Pr\{P \in \mathcal{D} \mid P_i \leq t\}$ is nondecreasing on the interval $(0, 1]$.
- Benjamini and Hochberg (BH): Given N hypothesis with corresponding p-values, reject the smallest $\widehat{k}$ p-values, where $\widehat{k}$ is chosen by:

$$\widehat{k} = \max\left\{k \mid P_{(k)} \leq \alpha \frac{k}{N}\right\}.$$

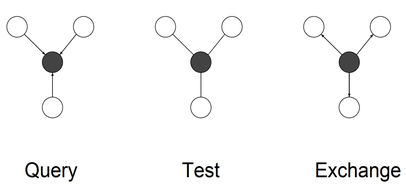
It guarantees FDR $\leq \alpha$ under independence, PRDS.

SINGLE-STEP QUTE ALGORITHM

Consider the following QuTE algorithm:

- **Query**: Each agent queries its neighbors and receives an ordered vector of p-values from each of their neighbors, and keeps track of the source of each p-value.
- **Test**: Let S_i be the set p-values now in possession of agent i. Agent i runs the BH procedure on its $|S_i|$ p-values, at an adjusted target FDR of $\alpha^i := \alpha\frac{|S_i|}{N}$.
- **Exchange**: All agents exchange their rejection decisions with their neighbors by sending back an ordered indicator vector, informing the neighbors to reject every hypothesis that has its indicator location set to unity.

In summary, an agent rejects a hypothesis whenever it was rejected by its own test, or by any of its neighbors.



MULTI-STEP QUTE ALGORITHM

With $c \geq 1$ rounds of communication, a node can access p-values of all nodes that are at a distance $\leq c$ away from it on the graph. After performing its local test, c rounds of exchanging information can once again propagate the results of its local test to nodes that are at distance $\leq c$ away. The fully dynamic algorithm is fully asynchronous — each node is constantly receiving new information, running local tests, and sharing its results.

QUTE GUARANTEES FDR CONTROL

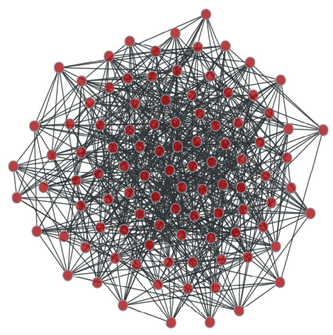
Theorem 1. *If the p-values are independent or positively dependent, then regardless of the graph topology, the multi-step QuTE algorithm achieves FDR control at level α for any $c \geq 1$.*

SIMULATIONS

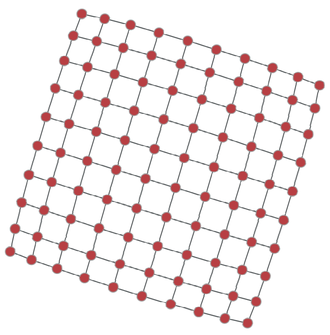
We generate the p-values as:

$$X \sim \mu + \mathcal{N}(0,1);\ \text{p-value} = 1 - \Phi(X),$$

where Φ is the standard Gaussian CDF, with $\mu = 0$ for nulls and $\mu > 0$ for alternatives. We consider the following two types of graphs:



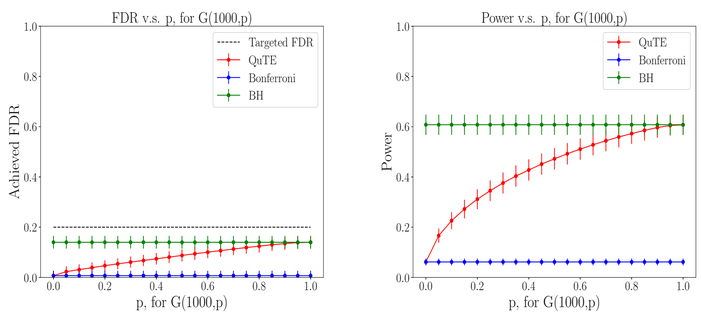
An example $G(n,p)$ graph



An example grid graph

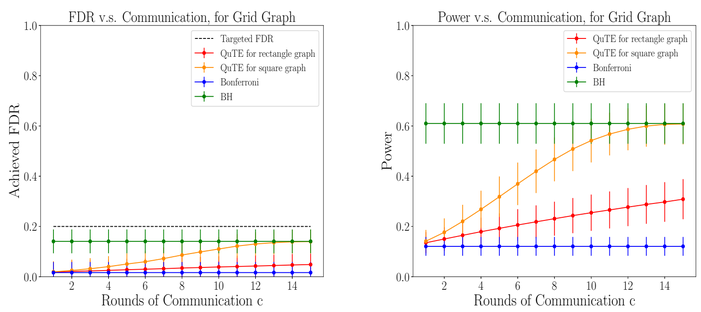
Erdös-Rényi random graph

A graph with n nodes is randomly generated, where an edge between any two distinct nodes is included in the graph with probability p independently.



Plots of FDR and power versus edge probability p for the QuTE, BH and Bonferroni procedures applied to the random graph model $G(1000, p)$, with target FDR $\alpha = 0.2$, signal level $\mu = 2$ and non-null frequency $\pi_1 = 0.3$.

Grid graph

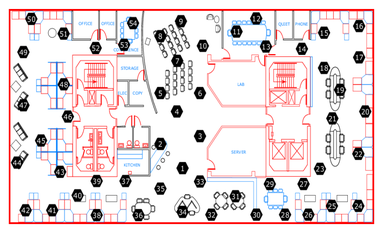


Plots of FDR and power versus rounds of communication with multi-step QuTE for a 16×16 grid graph, multi-step QuTE for a 2×128 grid graph, BH procedure, and Bonferroni procedure, with target FDR $\alpha = 0.2$, signal level $\mu = 2$ and the non-null frequency $\pi_1 = 0.3$.

AN ILLUSTRATIVE EXAMPLE ON REAL DATA

Dataset and preprocessing

Intel Lab dataset contains temperature data collected from 54 sensors deployed in the Intel Berkeley Research Lab across 36 days. Each pair of sensors has a communication probability. We threshold the probability and put an edge between two sensors if their communication probability with each other is larger than a certain threshold specified below.

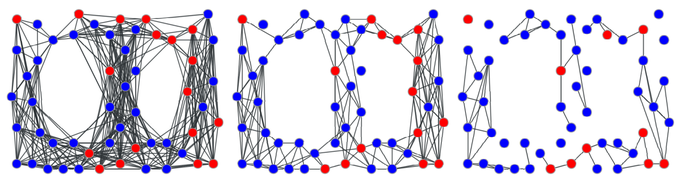


At each node, the null hypothesis states that the current temperature at the node is normal. Define the empirical distribution of the temperature of the first 500 samples at Sensor i as $\widehat{F}_{n,i}$. The p-value at node i is

$$p_i = 2\min(1 - \widehat{F}_{n,i}(t_i), \widehat{F}_{n,i}(t_i)),$$

Results

The graphs are specified by choosing the threshold to be $0.1, 0.3, 0.5$.



Threshold $\gamma = 0.1$ Threshold $\gamma = 0.3$ Threshold $\gamma = 0.5$

The nodes in red are rejected. Naturally, there are more rejections on average with increasing density of edges.

IMPORTANT EXTENSIONS

The multi-step QuTE algorithm is very robust, and the following extensions make it particularly practical.

Quantization

We designed a novel "quantized-BH" algorithm: Define the *p-rank* $R_i = \text{Ceiling}(p_i N/\alpha)$ and find

$$k^* = \max\{k : R_{(k)} \leq k\}.$$

In the Test stage of QuTE, we assume all unknown p-ranks equal $N + 1$ and run a local quantized-BH test on all N p-ranks at level α.

Time-varying Data

Suppose each sensor received more data as time passes. We define a false discovery to be a null that is wrongly rejected at *any* time. We can often construct an *always-valid p-value* at each node i, which is a stochastic process $\{p_i^t\}_{t \in \mathbb{R}}$ such that for any stopping time T, we have

$$\mathbb{P}\{p_i^T \leq x\} \leq x \ \text{ for any } \ x \in [0,1].$$

QuTE still guarantees FDR control if we replace static p-values by always-valid p-values.

Time-varying Edges

In the setting where the edges may arbitrarily drop packets, or sensors (such as drones) move around to change the graph topology, QuTE still guarantees FDR. Indeed, QuTE is fully asynchronous, and a node may receive information from different neighbors at different times.