Statistically Significant High Utility Itemset Mining — Random Database Generation

1 Notation

- $I \doteq \{i_1, \ldots, i_{|I|}\}$: set of all items I.
- $D \doteq \{T_1, \ldots, T_{|D|}\}$: quantitative transaction database D.
- $T_c \subseteq I$: transaction T_c that has a unique identifier c.
- $X \subseteq I$: itemset X.
- $p_D(i)$: external utility (i.e., price) of item i.
- $q(i, T_c)$: internal utility (i.e., quantity) of item i in transaction T_c.
- $q(T_c) \doteq \sum_{i \in T_c} q(i, T_c)$: total internal utility of a transaction T_c in database D.
- $q_D(i) \doteq \sum_{T_c \in D} q(i, T_c)$: total internal utility of item i in database D.
- $u(i, T_c) \doteq p_D(i) \cdot q(i, T_c)$: utility of item i in transaction T_c.
- $u(X, T_c) \doteq \sum_{i \in X} u(i, T_c)$: utility of an itemset X in transaction T_c.
- $u_D(X) \doteq \sum_{T_c \in D} u(X, T_c)$: utility of itemset X in database D.
- $TU(T_c) \doteq \sum_{i \in T_c} u(i, T_c) = u(T_c, T_c)$: transaction utility TU of transaction T_c.
- $s_D(X) \doteq |\{T_c \in D : X \subseteq T_c\}|$: the support of itemset X in database D.

2 Database Properties to Preserve

1. The external utility of each item i in database D, $p_D(i)$. This can be thought of as preserving the price of each item in the e-commerce store.
2. The total internal utility of each item i in database D, $q_D(i)$. This can be thought of as preserving the total purchased quantity of an item in the e-commerce store.

3. The distribution of transaction lengths in database D, $\{|T_c| : T_c \in D\}$. This can be thought of as preserving the number of unique items bought for each transaction in the e-commerce store.

4. The distribution of the transaction utilities in database D, $\{TU(T_c) : T_c \in D\}$. This can be thought of as preserving the amount of profit generated from each transaction in the e-commerce store.

5. The distribution of the transaction internal utilities in database D, $\{q(T_c) : T_c \in D\}$. This can be thought of as preserving the quantity of items bought for each transaction in the e-commerce store.

6. The distribution of internal utilities for each item across the transactions in database D, $\{q(i, T_c) : i \in T_c,\ T_c \in D\}\ \forall i \in I$. This can be thought of as preserving the purchased quantities of an item for each transaction in the e-commerce store.

3 Random Database Generation Methods

For the following methods, we will always preserve Prop. 1 since doing so is trivial.

1. For each item $i \in I$, randomly repartition $q_D(i)$ into $\leq |D|$ utilities. Randomly add i and each utility into a unique transaction in the new database D'. This method preserves Prop. 1 and Prop. 2.

2. Randomly select two transactions T_c and T_d with probability $\Pr(c) = \frac{|T_c|}{\sum_{j=1}^{|D|} |T_j|}$ and $\Pr(d) = \frac{|T_d|}{\sum_{j=1}^{|D|} |T_j|}$, respectively. Then, select items $i_k \in T_c$ and $i_l \in T_d$ uniformly at random within each transaction. If $i_k \notin T_d$ and $i_l \notin T_c$, then swap i_k and $q_D(i_k, T_c)$ with i_l and $q_D(i_l, T_d)$. Continue such execution for n steps. This method preserves Prop. 1 and Prop. 3.

3. Execute Method 2. Then, for each item $i \in I$, randomly repartition $q_D(i)$ into $s_D(\{i\})$ utilities. Place the repartitioned internal utilities back into the transactions that contain i. This method preserves Prop. 1, Prop. 2, and Prop. 3.

4. Execute Method 2. Then, for each item $i \in I$, construct a list $[q(i, T_c) : i \in T_c,\ T_c \in D]$. Permute the list and place the new ordering of internal utilities back into the transactions that contain i. This method preserves Prop. 1, Prop. 3, and Prop. 6.

5. Create a set of lists $U = \{L_j : L_j = [(i, T_c) : u(i, T_c) = j,\ i \in I,\ T_c \in D]\}$, where the lists in U contain item-transaction tuples of equal utility. Select a list $L_j \in U$ with probability $\Pr(j) = \frac{|L_j|^2}{\sum_{k=1}^{m} |L_k|^2}$, where $m = \max\{u(i, T_c) : i \in I,\ T_c \in D\}$, to ensure that item pairs are selected uniformly at random. Then, randomly select $(i_k, T_c), (i_l, T_d) \in L_j$ such

that $i_k \notin T_d$ and $i_l \notin T_c$. Next, swap (i_k, T_c) and (i_l, T_d). Continue such execution for n steps. This method preserves Prop. 1, Prop. 2, Prop. 3, Prop. 4, and Prop. 6.

6. Execute Method 5 but with internal utilities instead of utilities. This method preserves Prop. 1, Prop. 2, Prop. 3, Prop. 5, and Prop. 6.

4 Theorems

Theorem 4.1. *A random walk on a directed graph G converges to a stationary distribution $\bar{\pi}$, where*

$$\pi_v = \frac{d(v)}{|E|}.$$

Proof. Since $\sum_{v \in V} d(v) = |E|$, it follows that

$$\sum_{v \in V} \pi_v = \sum_{v \in V} \frac{d(v)}{|E|} = 1,$$

and $\bar{\pi}$ is a proper distribution over $v \in V$.

Let $\mathbf{P}$ be the transition probability matrix of the Markov chain. Let $N(v)$ represent the neighbors of v. The relation $\bar{\pi} = \bar{\pi}\mathbf{P}$ is equivalent to

$$\pi_v = \sum_{u \in N(v)} \frac{d(u)}{|E|} \frac{1}{d(u)} = \frac{d(v)}{|E|},$$

and the theorem follows. □

Theorem 4.2. *The stationary distribution of an irreducible aperiodic finite Markov chain is uniform if and only if its transition matrix is doubly stochastic.*

Corollary 4.2.1. *If the transition matrix of an irreducible aperiodic finite Markov chain is symmetric, then the stationary distribution of the Markov chain is uniform.*

Proof. Since the transition matrix is symmetric and its rows sum to 1, its columns also sum to 1. Thus, the transition matrix is doubly stochastic and so the stationary distribution is uniform. □

Lemma 4.3 (Metropolis-Hastings Algorithm)**.** *Suppose a Markov chain on a finite state space Ω is given by the transition matrix $\mathbf{Q}$ and a neighborhood structure $\{N(x) : x \in \Omega\}$. For all $x \in \Omega$, let π_x be the probability of state x in the stationary distribution of the desired Markov chain. Consider the Markov chain where*

$$P_{x,y} = \begin{cases} Q_{x,y} \min\left(\frac{\pi_y Q_{y,x}}{\pi_x Q_{x,y}}, 1\right) & \text{if } x \neq y \text{ and } y \in N(x). \\ 0 & \text{if } x \neq y \text{ and } y \notin N(x). \\ 1 - \sum_{y \neq x} P_{x,y} & \text{if } x = y. \end{cases}$$

Then, if this chain is irreducible and aperiodic, the stationary distribution is given by the probabilities π_x.

Proof. We show that the chain is time reversible and thus has a stationary distribution given by π_x. For any $x \neq y$, if $\pi_x Q_{x,y} \leq \pi_y Q_{y,x}$, then $P_{x,y} = Q_{x,y}$ and $P_{y,x} = \frac{\pi_x Q_{x,y}}{\pi_y}$. It follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. Similarly, if $\pi_x Q_{x,y} > \pi_y Q_{y,x}$, then $P_{x,y} = \frac{\pi_y Q_{y,x}}{\pi_x}$ and $P_{y,x} = Q_{y,x}$, and it follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. By Theorem 7.10 in the textbook, the stationary distribution is given by the values π_x. □