entropy

MDPI

Article

Recognizing Information Feature Variation: Message Importance Transfer Measure and Its Applications in Big Data

Rui She, Shanyun Liu and Pingyi Fan *

Department of Electronic Engineering, Tsinghua University, Beijing 30332, China;
sher15@mails.tsinghua.edu.cn (R.S.); liushany16@mails.tsinghua.edu.cn (S.L.)
* Correspondence: fpy@tsinghua.edu.cn; Tel.: +86-010-6279-6973

Received: 25 April 2018; Accepted: 12 May 2018; Published: 24 May 2018

Abstract: Information transfer that characterizes the information feature variation can have a crucial impact on big data analytics and processing. Actually, the measure for information transfer can reflect the system change from the statistics by using the variable distributions, similar to Kullback-Leibler (KL) divergence and Renyi divergence. Furthermore, to some degree, small probability events may carry the most important part of the total message in an information transfer of big data. Therefore, it is significant to propose an information transfer measure with respect to the message importance from the viewpoint of small probability events. In this paper, we present the message importance transfer measure (MITM) and analyze its performance and applications in three aspects. First, we discuss the robustness of MITM by using it to measuring information distance. Then, we present a message importance transfer capacity by resorting to the MITM and give an upper bound for the information transfer process with disturbance. Finally, we apply the MITM to discuss the queue length selection, which is the fundamental problem of caching operation on mobile edge computing.

Keywords: information transfer measure; small probability events; big data analysis and processing; mobile edge computing (MEC); queue theory

1. Introduction

In recent years, due to the exploding amount of data, computing complexity for data processing is growing rapidly. In particular, cloud data center traffic will jump up to one order of magnitude by 2020 [1,2]. To some degree, the reason for this phenomenon seems to be that more and more mobile devices such as smartphones, tablets or mobile Internet of things (IoT) devices are utilized and the growing services of clouds are provided. In this context, it is necessary to dig out the valuable information from the collected data. On one hand, computation technologies including cloud computing and mobile edge computing (MEC) are needed for big data processing. On the other hand, it is essential to develop more efficient technologies for big data analysis and mining, such as distributed parallel computing, machine learning, deep learning, and neural networks, etc. [3–6].

As for data mining, the small probability events usually attract much more attention than the large probability ones [7–10]. In other words, there exits higher practical value in the rarity of small probability events. For example, in the anti-terrorist scenario, we just focus on a few illegal and dangerous people [11]. Moreover, as for the synthetic identification (ID) detection, only a small number of artificial identities for financial frauds should be paid more attention to [12]. In fact, it is challenging and significant to measure and mine small probability events.

According to rate-distortion theory, it is rational for us to regard small probability events detection as a clustering problem [13,14]. By using popular clustering principles (e.g., minimum within-cluster distance, maximum inter-cluster distance, and minimum compressing distortion), some efficient clustering approaches were proposed to detect small probability events. Specifically, a graph-based rare category detection and time-flexible rare category detection were presented based on the global similarity matrix and time-evolving of graphs, respectively [15,16]. Actually, these algorithms were proposed by resorting to traditional information measures and theory, which are considered from the viewpoint of typical events, which are the large probability events.

1.1. Information Transfer Measures Based on Message Importance

1.1.1. Review of Message Importance Measure

In information theory, there are two fundamental measures, Shannon entropy and Renyi entropy, which have a vital impact on wireless communication, estimation theory, signal processing and pattern recognition etc. Nevertheless, they are not applicable to mining small probability events hidden in big data.

To do this, a new information measure named message importance measure (MIM) is proposed from the perspective of big data [17]. To simplify the form of MIM, we shall introduce the definition of MIM as follows.

Definition 1. *For a continuous probability distribution $f(x)$ with respect to the variable X in a given interval S_x, the differential message importance measure (DMIM) focusing on the small probability events is defined as*

$$L(f(x)) = \int_{S_x} f(x)e^{-f(x)}dx, \qquad x \in S_x. \tag{1}$$

Furthermore, for the discrete probability P=$\{p(x_1), p(x_2), \ldots, p(x_n)\}$, the relative message importance measure (RMIM) is given by

$$L(P) = \sum_{x_i} p(x_i)e^{-p(x_i)}. \tag{2}$$

By resorting to the exponential form, the MIM can amplify small probability elements much more than Shannon entropy and Renyi entropy, which include the logarithm operator or polynomial operator respectively. Actually, this highlights the significance of small probability events in information measure and theory. In addition, a series of postulates are investigated to characterize Shannon entropy and Renyi entropy. Particularly, Fadeev's postulates are well-known to describe the information measures, which consist of four postulates [18]. In this case, in terms of two independent random distributions P and Q, there exists a weaker postulate for Renyi entropy than that for Shannon entropy, as follows

$$H(PQ) = H(P) + H(Q), \tag{3}$$

where the function $H(\cdot)$ denotes a kind of information measure. Similarly, there exists a weaker postulate for the MIM than that for Renyi entropy, namely

$$H(PQ) \leq H(P) + H(Q). \tag{4}$$

Consequently, from the viewpoint of generalized Fadeev's postulates, we can regard the MIM as a reasonable information measure similar to Shannon entropy and Renyi entropy.

1.1.2. Message Importance Transfer Measure

As for an information transfer process, we construct such a model that the original probability distribution P and the final one Q in the transfer process satisfies the Lipschitz condition as follows,

$$|H(P) - H(Q)| \leq \lambda \|P - Q\|_1, \tag{5}$$

where $H(\cdot)$ is the corresponding information measure function; $\lambda > 0$ is the Lipschitz constant; $\|\cdot\|_1$ denotes the l_1-norm measure.

Here, we shall analyze and measure the information transfer process mentioned in Equation (5) from the perspective of the message importance. In fact, it is a significant problem for us to measure the message importance variation in big data analytics. According to Definition 1, it is available to regard the DMIM or RMIM as an element to measure the message importance distance which can be also used in the discussion of information transfer processes. Then, an information transfer measure focusing on the message importance are proposed as follows.

Definition 2. *For two probability distributions $g(x)$ and $f(x)$ with respect to the variable X in a given interval S_x, the message importance transfer measure (MITM) is defined as*

$$\begin{aligned} &D_I(g(x)||f(x)) \\ &= L(g(x)) - L(f(x)) \\ &= \int_{S_x} \left(g(x)e^{-g(x)} - f(x)e^{-f(x)} \right) dx, x \in S_x. \end{aligned} \tag{6}$$

Furthermore, in terms of the two discrete probability $Q = \{q(x_1), q(x_2), \dots, q(x_n)\}$ and $P = \{p(x_1), p(x_2), \dots, p(x_n)\}$, the MITM can be written as

$$D_I(Q||P) = \sum_{x_i} \{q(x_i)e^{-q(x_i)} - p(x_i)e^{-p(x_i)}\}. \tag{7}$$

Note that Definition 2 characterizes a kind of relationship between two distributions from the perspective of information theory. In fact, this is a reasonable information measure that focuses on the effects of small probability elements regarded as message importance for two end-to-end distributions. On one hand, the MITM provides a tool to reflect the change of message importance in the whole transfer process. On the other hand, it also reveals the entire information feature variation of two end-to-end distributions, which we can use as a promising tool in the data mining.

1.2. Related Works for Information Measures in Big Data

There exist a variety of different information measures handling the problem of distributions, which can play a crucial role in many applications involved with artificial intelligence as well as big data analysis and processing.

As typical information measures, Shannon entropy and Renyi entropy are applicable to texture classification, intrinsic dimension estimation [19]. As well, the relative entropy, a kind of K-L divergence, is suitable for outlier detection [20] and functional magnetic resonance imaging (FMRI) data processing [21]. Moreover, the MIM and non-parametric message importance measure (NMIM) both focusing on the small probability events, have been proven effective in anomaly detection [17,22,23]. What is more, information divergences such as message importance (M-I) divergence can be applicable to extending methods of machine learning by using distributions and their relationship as features [24].

In addition, some information measures are proposed to reveal the correlation of message during the information transfer process. For example, the directed information [25–28] and Schreiber's transfer entropy [29] are commonly applied to infer the causality structure and characterize the information transfer process. Moreover, referring to the idea from dynamical system theory, new information transfer measures are proposed to indicate the causality between states and control the systems [30–32].

However, in spite of numerous kinds of information measures, few works focus on how to characterize the information transfer from the perspective of message importance in big data. To this end, a new information measure different from the above is introduced.

1.3. Organization

We organize the rest of this paper as follows. In Section 2, we investigate the variation of message importance in the information transfer process by using MITM. In Section 3, we introduce the message importance transfer capacity measured by the MITM to describe the information transfer system with additive disturbance. In Section 4, the MITM and the KL divergence are used to guide the queue length selection for MEC from the viewpoint of the queue theory. Moreover, we also present some simulations to validate our theoretical results. Finally, we conclude in Section 5.

2. The Information Distance for Message Importance Variation

We now investigate the variation of message importance between two distributions by using an information transfer measure. This characterizes the information distance from the perspective of message importance, which can also reflect the robustness of the information transfer measure.

Consider an observation model, $\mathcal{P}_{g_0|f_0}$: $f_0(x) \to g_0(x)$, namely an information transfer map for the variable X from one distribution $f_0(x)$ to the other distribution $g_0(x)$. In fact, it turns out to be not easy to cope with the two distributions. Instead, considering the similar way in [33], the relationship between $f_0(x)$ and $g_0(x)$ is given by

$$g_0(x) = f_0(x) + \epsilon f_0^{\alpha}(x)u(x), \tag{8}$$

and the constraint condition satisfies

$$\int_{S_x} \epsilon f_0^{\alpha}(x)u(x)dx = 0, \tag{9}$$

where ϵ and α are two positive adjustable coefficients, as well as $u(x)$ is a perturbation function of the variable X in the interval S_x.

Then, we discuss the information distance of message importance measured by the MITM in the Definition 2. This characterizes the difference between the origin and the destination of the information transfer from the viewpoint of message importance. By using the model $\mathcal{P}_{g_0|f_0}$: $f_0(x) \to g_0(x)$ mentioned above, the end-to-end MITM is investigated in the information transfer process as follows.

Proposition 1. *For two probability distributions $g_0(x)$ and $f_0(x)$ whose relationship satisfies the conditions Equations (8) and (9), the MITM is given by*

$$\begin{aligned} & D_I(g_0(x)||f_0(x)) \\ & = \int_{S_x} \left\{ g_0(x)e^{-g_0(x)} - f_0(x)e^{-f_0(x)} \right\} dx \\ & = \epsilon \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{i!} \int_{S_x} f_0^{i+\alpha}(x)u(x)dx \\ & + \frac{\epsilon^2}{2} \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{(i-1)!} \int_{S_x} f_0^{i-1+2\alpha}(x)u^2(x)dx + o(\epsilon^2), \end{aligned} \tag{10}$$

where ϵ and α are parameters, $u(x)$ denotes a function of the variable X, and $|D_I(g_0(x)||f_0(x))| \leq \int_{S_x} |\epsilon f_0^{\alpha}(x)u(x)|dx$ that satisfies the constraint Equation (5).

Proof of Proposition 1. According to the Binomial theorem, it is not difficult to see that

$$\begin{aligned} & g_0^i(x) - f_0^i(x) \\ & = [f_0(x) + \epsilon f_0^{\alpha}(x)u(x)]^i - f_0^i(x) \\ & = \sum_{r=1}^{i} C_i^r f_0^{i-r}(x)[\epsilon f_0^{\alpha}(x)u(x)]^r. \end{aligned} \tag{11}$$

Then, by using Taylor series expansion of e^x, it is readily seen that

$$\begin{aligned} & e^{-g_0(x)} - e^{-f_0(x)} \\ & = \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} [g_0^i(x) - f_0^i(x)] \\ & = \epsilon \sum_{i=1}^{\infty} \frac{(-1)^i}{(i-1)!} f_0^{(i-1+\alpha)}(x) u(x) + \frac{\epsilon^2}{2} \sum_{i=2}^{\infty} \frac{(-1)^i}{(i-2)!} f_0^{i-2+2\alpha}(x) u^2(x) + o(\epsilon^2). \end{aligned} \tag{12}$$

Therefore, by substituting Equation (12) into Equation (6), the proof of the proposition can be readily completed. □

Furthermore, it is not difficult to gain the MITM between the two different distributions $g_1^{(u)}$ and $g_2^{(u)}$ based on the same reference distribution $f_0(x)$ as follows

$$\begin{aligned} & D_I(g_1^{(u)}(x) || g_2^{(u)}(x)) \\ & = [L(g_1^{(u)}(x)) - L(f_0(x))] - [L(g_2^{(u)}(x)) - L(f_0(x))] \\ & = \epsilon \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{i!} \int_{S_x} f_0^{i+\alpha}(x) [u_1(x) - u_2(x)] dx \\ & + \frac{\epsilon^2}{2} \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{(i-1)!} \int_{S_x} f_0^{i-1+2\alpha}(x) [u_1^2(x) - u_2^2(x)] dx + o(\epsilon^2), \end{aligned} \tag{13}$$

where

$$g_1^{(u)}(x) = f_0(x) + \epsilon f_0^{\alpha}(x) u_1(x), \quad \forall x \in S_x, \tag{14a}$$

$$g_2^{(u)}(x) = f_0(x) + \epsilon f_0^{\alpha}(x) u_2(x), \quad \forall x \in S_x, \tag{14b}$$

in which the ϵ and α are parameters, $u_1(x)$ and $u_2(x)$ denote functions of the variable X in the interval S_x, and $|D_I(g_1(x)||g_2(x))| \leq \int_{S_x} |\epsilon f_0^{\alpha}(x)\{u_1(x) - u_2(x)\}| dx$.

Similarly, it is available for the discrete probability distributions to have the same form of MITM as that mentioned in the Proposition 1. In particular, for two distributions $Q_0 = \{q_0(x_1), q_0(x_2), \ldots, q_0(x_n)\}$ and $P_0 = \{p_0(x_1),\ p_0(x_2), \ldots, p_0(x_n)\}$, it is easy to see that if the relationship between Q_0 and P_0 satisfies

$$q_0(x_i) = p_0(x_i) + \epsilon p_0^{\alpha}(x_i) \tilde{u}(x_i), \tag{15}$$

with the constraint condition $\sum_{x_i} p_0^{\alpha}(x_i) \tilde{u}(x_i) = 0$, we will have

$$\begin{aligned} & D_I(Q_0 || P_0) \\ & = \epsilon \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{i!} \sum_{x_i} p_0^{i+\alpha}(x_i) \tilde{u}(x_i) + \frac{\epsilon^2}{2} \sum_{i=1}^{\infty} \frac{(-1)^i (i+1)}{(i-1)!} \sum_{x_i} p_0^{i-1+2\alpha}(x_0) \tilde{u}^2(x_0) + o(\epsilon^2), \end{aligned} \tag{16}$$

where ϵ and α are adjustable coefficients, and $\tilde{u}(x_i)$ is a perturbation function of the variable X. Moreover, it is not difficult to gain the discrete form of Equation (13) in the same way as above.

Remark 1. *By resorting to the information distance measured by the MITM, the message importance distinction between two different distributions can be characterized. In the observation model mentioned in Equation (8), it is apparent that the parameter ϵ dominates the information distance when the perturbation function is finite and the parameter $\alpha < \infty$. Furthermore, the MITM is convergent with the order of $O(\epsilon)$ in the case of small parameter ϵ. Actually, it provides a way to apply MITM to measure the message importance variation in an information transfer process.*

3. Message Importance Transfer Capacity

In this section, we shall utilize the MITM to analyze the information transfer processing shown in Figure 1. To this end, we propose the message importance transfer capacity based on the MITM as follows.

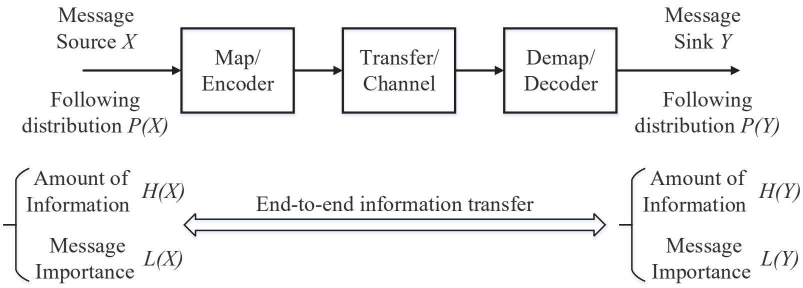


Figure 1. Information transfer system model.

Definition 3. *Assume that there exists an information transfer process*

$$\{X, p(y|x), Y\}, \tag{17}$$

where the $p(y|x)$ is a probability distribution matrix characterizing the information transfer from the variable X to Y. The message importance transfer capacity is defined as

$$C = \max_{p(x)}\{L(Y) - L(Y|X)\}, \tag{18}$$

where $p(y) = \int_{S_x} p(x)p(y|x)dx$, $L(Y) = \int_{S_y} p(y)e^{-p(y)}dy$, and $L(Y|X) = \int_{S_x}\int_{S_y} p(xy)e^{-p(y|x)}dxdy$ with the constraint $|L(Y) - L(Y|X)| \leq \lambda\|p(y) - p(y|x)\|_1$.

Then, we discuss some specific information transfer scenarios to have an insight into the applications of message importance transfer capacity, as follows.

3.1. Binary Symmetric Information Transfer Matrix

Consider the binary symmetric information transfer matrix, in which the original variables are complemented with the transfer probability. In particular, the rows of the probability matrix are permutations of each other and so are columns which can be seen in the following proposition.

Proposition 2. *Assume an information transfer process $\{X, p(y|x), Y\}$, whose the information transfer matrix is described as*

$$p(y|x) = \begin{bmatrix} 1-\beta & \beta \\ \beta & 1-\beta \end{bmatrix}, \tag{19}$$

which implies that the variable X and Y both follow binary distributions. In this case, we have the message importance transfer capacity as follows

$$C(\beta) = e^{-\frac{1}{2}} - L(\beta), \tag{20}$$

where $L(\beta) = \beta e^{-\beta} + (1-\beta)e^{-(1-\beta)}$ with $0 < \beta < 1$, and $|C(\beta)| \leq \lambda\|p(y) - p(y|x)\|_1$ with $\lambda \geq \frac{e^{-\frac{1}{2}} - \beta e^{-\beta} + (1-\beta)e^{-(1-\beta)}}{|1-2\beta|}$.

Proof of Proposition 2. Assume that the distribution of variable X is a binary distribution $(p, 1-p)$. As well, it is readily seen that

$$\begin{aligned} L(Y|X) &= \sum_{i=1}^{2} p(x_i) \sum_{y} p(y|x_i) e^{-p(y|x_i)} \\ &= \sum_{y} p(y|x_i) e^{-p(y|x_i)} \\ &= \beta e^{-\beta} + (1-\beta) e^{-(1-\beta)} = L(\beta). \end{aligned} \tag{21}$$

Moreover, according to the definition of C in Equation (18), we have

$$\begin{aligned} &C(p,\beta) \\ &= \max_{p} \left\{ [p + \beta(1-2p)] e^{-[p+\beta(1-2p)]} + [(1-p) + \beta(2p-1)] e^{-[(1-p)+\beta(2p-1)]} \right\} - L(\beta). \end{aligned} \tag{22}$$

Then, it is not difficult to see that

$$\begin{aligned} &\frac{\partial C(p,\beta)}{\partial p} \\ &= (1-2\beta) \left\{ [1 - p - \beta(1-2p)] e^{-[p+\beta(1-2p)]} - [1 - (1-p) - \beta(2p-1)] e^{-[(1-p)+\beta(2p-1)]} \right\}. \end{aligned} \tag{23}$$

According to the monotonically decreasing of $\frac{\partial C(p,\beta)}{\partial p}$ for $p \in [0,1]$, it is readily seen that $p = \frac{1}{2}$ is the only solution for $\frac{\partial C(p,\beta)}{\partial p} = 0$. Therefore, by substituting $p = \frac{1}{2}$ into $C(p,\beta)$, the proposition is testified. □

Remark 2. *In light of Proposition 2, on one hand, when $\beta = 1/2$, in other words, there is just random information transfer process, we will obtain the lower bound of the message importance transfer capacity that is $C(\beta) = 0$. On the other hand, when $\beta = 0$, namely, the information transfer process is definite, we will gain the maximum message importance transfer capacity.*

3.2. Binary Erasure Information Transfer Matrix

The binary erasure information transfer matrix is similar to the binary symmetric one, however, in the former a part of information is lost rather than corrupted. In other words, a fraction of information is erased. In this case, the message importance transfer capacity is discussed as follows.

Proposition 3. *Consider an information transfer process $\{X, p(y|x), Y\}$, in which the information transfer matrix is described as*

$$p(y|x) = \begin{bmatrix} 1-\beta & 0 & \beta \\ 0 & 1-\beta & \beta \end{bmatrix}, \tag{24}$$

which indicates that X follows the binary distribution and Y follows the 3-ary distribution. Then, we have

$$C(\beta) = (1-\beta) e^{-\frac{1}{2}(1-\beta)} + \beta e^{-\beta} - L(\beta), \tag{25}$$

where $L(\beta) = \beta e^{-\beta} + (1-\beta) e^{-(1-\beta)}$ with $0 < \beta < 1$ and $|C(\beta)| \le \lambda \| p(y) - p(y|x) \|_1$ with $\lambda \ge e^{-\frac{1}{2}(1-\beta)} - e^{-(1-\beta)}$.

Proof of Proposition 3. Assume the distribution of variable X is $(p, 1-p)$. As well, according to the binary erasure information transfer matrix, it is not difficult to see that

$$C(p,\beta) = \max_{p} \left\{ p(1-\beta) e^{-p(1-\beta)} + \beta e^{-\beta} + (1-p)(1-\beta) e^{-(1-p)(1-\beta)} \right\} - L(\beta), \tag{26}$$

where $L(\beta) = \beta e^{-\beta} + (1-\beta)e^{-(1-\beta)}$. Then, we have

$$\frac{\partial C(p,\beta)}{\partial p} = (1-\beta)\Big\{[1-(1-\beta)p]e^{-(1-\beta)p} - [1-(1-\beta)(1-p)]e^{-(1-\beta)(1-p)}\Big\}. \tag{27}$$

Due to the monotonically decreasing of $\frac{\partial C(p,\beta)}{\partial p}$ for $p \in [0,1]$, it is readily seen that $p = 1/2$ is the only solution for $\frac{\partial C(p,\beta)}{\partial p} = 0$. Thus, by substituting $p = 1/2$ into Equation (26), the proposition is readily verified. □

3.3. Strongly Symmetric Information Transfer Matrix

In terms of the strongly symmetric information transfer matrix, it can be regarded as an extension of the binary symmetric one. The message information transfer capacity of the former is also analogous to the that of the latter, which is discussed as follows.

Proposition 4. *Assume an information transfer process with the strongly symmetric transfer matrix as follows*

$$p(y|x) = \begin{bmatrix} 1-\beta & \frac{\beta}{K-1} & \cdots & \frac{\beta}{K-1} \\ \frac{\beta}{K-1} & 1-\beta & \cdots & \frac{\beta}{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta}{K-1} & \cdots & \frac{\beta}{K-1} & 1-\beta \end{bmatrix}, \tag{28}$$

which implies that the variable X and Y both obey K-ary distribution. We have

$$C(\beta) = e^{-\frac{1}{K}} - \{(1-\beta)e^{-(1-\beta)} + \beta e^{-\frac{\beta}{K-1}}\}, \tag{29}$$

where the parameter $\beta \in (0,1)$ *and* $|C(\beta)| \leq \lambda \|p(y) - p(y|x)\|_1$ *with* $\lambda \geq \frac{e^{-1/K} - (1-\beta)e^{-(1-\beta)} - \beta e^{-\beta/K-1}}{2|1-\beta-1/K|}$.

Proof of Proposition 4. Assume the probability distribution of variable X is $\{p(x_1), p(x_2), \ldots, p(x_K)\}$. As for the strongly symmetric transfer matrix, when the probabilities of x_i are equal, that is, $p(x_1) = p(x_2) = \ldots = p(x_K) = 1/K$, we will have

$$\begin{aligned} p(y_j) &= \sum_{i=1}^{K} p(x_i, y_j) = \sum_{i=1}^{K} p(x_i)p(y_j|x_i) \\ &= \frac{1}{K}\sum_{i=1}^{K} p(y_j|x_i) = \frac{1}{K}, \end{aligned} \tag{30}$$

which indicates that the probabilities of y_j $(j = 1, 2, \ldots, K)$ are equal.

In addition, on account of the information transfer matrix, it is easy to see that

$$\begin{aligned} L(Y|X) &= \sum_{i=1}^{2} p(x_i) \sum_{y_j} p(y_j|x_i)e^{-p(y_j|x_i)} \\ &= \sum_{y_j} p(y_j|x_i)e^{-p(y_j|x_i)} \\ &= \beta e^{-\frac{\beta}{K-1}} + (1-\beta)e^{-(1-\beta)}. \end{aligned} \tag{31}$$

What is more, according to the definition of message importance transfer capacity in Equation (18), it is readily seen that

$$C(\beta) = \max_{p(x)}\{L(Y)\} - [\beta e^{-\frac{\beta}{K-1}} + (1-\beta)e^{-(1-\beta)}], \tag{32}$$

where $L(Y) = \sum_{y_j} p(y_j)e^{-p(y_j)}$.

Then, by using Lagrange multiplier method, we have

$$G(p(y_j), \lambda_0) = \sum_{y_j} p(y_j)e^{-p(y_j)} + \lambda_0 \big[\sum_{y_j} p(y_j) - 1 \big]. \tag{33}$$

By setting $\frac{\partial G(p(y_j),\lambda_0)}{\partial p(y_j)} = 0$ and $\frac{\partial G(p(y_j),\lambda_0)}{\partial \lambda_0} = 0$, it can be readily verified that the extreme value of $\sum_{y_j} p(y_j)e^{-p(y_j)}$ is achieved by the solution $p(y_1) = p(y_2) = \ldots = p(y_K) = 1/K$.

In light of $\frac{\partial^2 G(p(y_j),\lambda_0)}{\partial p^2(y_j)} < 0$ with respect to $p(y_j) \in [0,1]$, it is readily seen that when the variable X follows the uniform distribution which leads to the uniform distribution for variable Y, we will gain the message importance transfer capacity $C(\beta)$. Then, it is easy for us to complete the proof of the proposition. □

3.4. Continuous Case for the Message Importance Transfer Capacity

By using the MITM as a measuring tool, the information transfer process in the continuous case is investigated. Considering the information transfer process described as Equation (17), it is significant to clarify the effect of the continuous disturbance on the message importance transfer capacity.

Theorem 1. *Assume that there exists an information transfer process between the variable X and Y, denoted by $\{X, p(y|x), Y\}$, where $E[X] = 0$, $E[X^2] = P_s$, $Y = X + Z$. The variable Z denotes an independent memoryless additive disturbance, whose mean and variance satisfy that $E[Z] = \mu$ and $E[(Z-\mu)^2] = \sigma^2$, respectively. Then, we adopt the MITM to measure the message importance transfer capacity as*

$$\begin{aligned} C(P_s) &= \max_{p(x)} D_I(Y||Z) \\ &= \max_{p(x)} \{L(Y)\} - L(Z) \\ &= \max_{p(x)} \Big\{ \int_{-\infty}^{+\infty} p(y)e^{-p(y)}dy \Big\} - \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \int_{-\infty}^{\infty} p^{j+1}(z)dz, \end{aligned} \tag{34a}$$

$$s.t. \quad E[Y^2] = P_s + P_N, \tag{34b}$$

where $P_N = \mu^2 + \sigma^2$, $p(y) = \int_{S_x} p(x)p(y|x)dx$ with the constraint $|L(Y) - L(Z)| \leq \lambda \|p(y) - p(z)\|_1$ ($\lambda > 0$ is the Lipschitz constant), and $L(\cdot)$ is the MIM operator. That is, the variance of X makes more effect on the constraint of the message importance transfer capacity.

Proof of Theorem 1. According to Equation (17), we have

$$\begin{cases} x(x,y) = x \\ z(x,y) = y - x. \end{cases} \tag{35}$$

Then, it is not difficult to see that

$$p(xy) = p(xz)|J(\frac{xz}{xy})| = p(xz) \begin{vmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} \\ \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{vmatrix} = p(xz). \tag{36}$$

Moreover, by virtue of the independence of X and Z, we have

$$p(x)p(y|x) = p(x)p(z), \tag{37}$$

which indicates that

$$p(y|x) = p(z). \tag{38}$$

Then, we have

$$\begin{aligned} L(Y|X) &= \int_x \int_y p(x)p(y|x)e^{-p(y|x)}dxdy \\ &= \int_x \int_z p(x)p(z)e^{-p(z)}dxdz \\ &= \int_x p(x)\{\int_z p(z)e^{-p(z)}dz\}dx = L(Z). \end{aligned} \tag{39}$$

Consequently, in terms of the Definition 3, it is readily seen that $L(Y) - L(Y|X)$ can be written as $L(Y) - L(Z)$, which testifies Equation (34a).

Furthermore, according to the fact that $E[Y^2] = E[(X+Z)^2] = E[X^2] + E[Z^2] = P_s + P_N$, we have the constraint condition Equation (34b). As well, by substituting the definition of MITM into Equation (34a), the Theorem 1 is proved. □

Remark 3. *For the message importance transfer capacity with an additive disturbance, it is worth noting that the distribution of the transferred variable Y with the constrained variance may have a significant impact on the practical applications. In practice, the variance can be regarded as the power of signals. Consequently, the message importance transfer capacity mentioned in Theorem 1 can be used to guide the signal transfer process with additive disturbance, if the system does not have relatively large change.*

Corollary 1. *Consider an information transfer process $\{X, p(y|x), Y\}$, where $Y = X + Z$ and the variable Z denotes an independent Gaussian disturbance with $E[Z] = \mu_z$ and $E[Z^2] = \sigma_z^2$. Assume that the variable X follows a Gaussian mixture model as*

$$P_X(x) = \frac{1}{N}\sum_{k=1}^{k=N} \phi(x|\mu_k, \sigma_k^2), \tag{40}$$

where μ_k and σ_k^2 are the means and the variances of independent Gaussian distributions, in other words, $\phi(x|\mu_k, \sigma_k^2) = 1/(\sqrt{2\pi}\sigma_k)\exp\{-(x-\mu_k)^2/(2\sigma_k^2)\}$. In this case, the message importance transfer capacity $C(\mu_x, \sigma_x^2)$ with the constraint $|C(\mu_x, \sigma_x^2)| \le \lambda\|P_Y(y) - P_Z(z)\|_1$, is

$$\begin{aligned} C(\mu_x, \sigma_x^2) &= \max_{P_X(x)} D_I(Y||Z) \\ &\doteq \frac{1}{2\sqrt{\pi\sigma_z^2}} - \frac{1}{2N^2\sqrt{\pi(\Theta+\sigma_z^2)}}\sum_{i=1}^{N}\sum_{j=1}^{N} e^{-\frac{(\mu_i-\mu_j)^2}{4(\Theta+\sigma_z^2)}}, \end{aligned} \tag{41}$$

where $\Theta = \frac{1}{N}\sum_{k=1}^{N}\sigma_k^2$. In particular, the parameters σ_k^2 can be controlled by the parameters σ_x^2, μ_x and μ_k in a system, where the μ_x and σ_x^2 are the mean and variance of the variable X, which are given by

$$\mu_x = \frac{1}{N}\sum_{k=1}^{N}\mu_k, \tag{42a}$$

$$\sigma_x^2 = \frac{1}{N}\sum_{k=1}^{N}(\sigma_k^2 + \mu_k^2) - \left(\frac{1}{N}\sum_{k=1}^{N}\mu_k\right)^2. \tag{42b}$$

Proof of Corollary 1. As for the Gaussian variable Z satisfying $E[Z] = \mu_z$ and $E[Z^2] = \sigma_z^2$, the DMIM is given by

$$\begin{aligned} L(Z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{(z-\mu_z)^2}{2\sigma_z^2}} e^{-\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(z-\mu_z)^2}{2\sigma_z^2}}} dz \\ &= \sum_{i=0}^{\infty} \frac{(-1)^i}{i!(\sqrt{2\pi\sigma_z^2})^{i+1}} \frac{1}{2}\sqrt{\frac{\pi}{\alpha_0}} e^{\frac{\beta_0^2-\alpha_0\gamma_0}{\alpha_0}} \cdot erf\left(\sqrt{\alpha_0}z + \frac{\beta_0}{\sqrt{\alpha_0}}\right)\Big|_{z=-\infty}^{z=\infty}, \end{aligned} \tag{43}$$

where the $erf(\cdot)$ is the error function, namely,

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt, \tag{44}$$

and the parameters α_0, β_0 and γ_0 satisfy

$$\alpha_0 = \frac{i+1}{2\sigma_z^2}, \tag{45a}$$

$$\beta_0 = -\frac{(i+1)\mu_z}{2\sigma_z^2}, \tag{45b}$$

$$\gamma_0 = \frac{(i+1)\mu_z^2}{2\sigma_z^2}. \tag{45c}$$

Then, it is readily seen that

$$L(Z) = \sum_{i=0}^{\infty} \frac{(-1)^i}{i!\sqrt{i+1}(\sqrt{2\pi\sigma_z^2})^i}, \tag{46}$$

which can be approximated by

$$L(Z) \doteq 1 - \frac{1}{2\sqrt{\pi\sigma_z^2}}. \tag{47}$$

In addition, according to $Y = X + Z$ (with the independent X and Z), it is readily seen that the variable Y also follows a Gaussian mixture model as

$$\begin{aligned} &P_Y(y) \\ &= \int_{-\infty}^{\infty} P_X(x) P_Z(y-x) dx \\ &= \frac{1}{N} \sum_{k=1}^{k=N} \int_{-\infty}^{\infty} \phi(x|\mu_k, \sigma_k) \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{(y-x-\mu_z)^2}{2\sigma_z^2}} dx \\ &= \frac{1}{N} \sum_{k=1}^{k=N} \frac{1}{\sqrt{2\pi(\sigma_k^2+\sigma_z^2)}} e^{-\frac{(y-\mu_k-\mu_z)^2}{2(\sigma_z^2+\sigma_k^2)}} \\ &= \frac{1}{N} \sum_{k=1}^{k=N} \phi(y|\tilde{\mu}_k, \tilde{\sigma}_k^2), \end{aligned} \tag{48}$$

where $\tilde{\mu}_k = \mu_k + \mu_z$ and $\tilde{\sigma}_k^2 = \sigma_k^2 + \sigma_z^2$ $(k = 1, 2, \ldots, N)$.

By using of Taylor series extension, we have the DMIM of variable Y as follows

$$\begin{aligned} L(Y) &= \int_{-\infty}^{\infty} P_Y(y) e^{-P_Y(y)} dy \\ &= \int_{-\infty}^{\infty} P_Y(y)[1 - P_Y(y) + O(P_Y^2(y))] dy \\ &\doteq 1 - \int_{-\infty}^{\infty} P_Y^2(y) dy. \end{aligned} \tag{49}$$

Then, according to Equation (48), it is readily seen that

$$\begin{aligned} L(Y) &\doteq 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{\infty} \phi(y|\tilde{\mu}_i, \tilde{\sigma}_i^2) \phi(y|\tilde{\mu}_j, \tilde{\sigma}_j^2) dy \\ &= 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{2\pi\sqrt{\sigma_i^2\sigma_j^2}} \left\{ \frac{1}{2} \sqrt{\frac{\pi}{\alpha_1}} e^{\frac{\beta_1^2 - \alpha_1\gamma_1}{\alpha_1}} \cdot erf\left(\sqrt{\alpha_1} y + \frac{\beta_1}{\sqrt{\alpha_1}} \right) \Big|_{y=-\infty}^{y=\infty} \right\}, \end{aligned} \tag{50}$$

where $\tilde{\mu}_k = \mu_k + \mu_z$ and $\tilde{\sigma}_k^2 = \sigma_k^2 + \sigma_z^2$ $(k = 1, 2, \ldots, N)$, the parameters α_1, β_1 and γ_1 are

$$\alpha_1 = \frac{\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2}{2\tilde{\sigma}_i^2\tilde{\sigma}_j^2}, \tag{51a}$$

$$\beta_1 = -\frac{\tilde{\mu}_i\tilde{\sigma}_j^2 + \tilde{\mu}_j\tilde{\sigma}_i^2}{2\tilde{\sigma}_i^2\tilde{\sigma}_j^2}, \tag{51b}$$

$$\gamma_1 = \frac{\tilde{\mu}_i^2\tilde{\sigma}_j^2 + \tilde{\mu}_j^2\tilde{\sigma}_i^2}{2\tilde{\sigma}_i^2\tilde{\sigma}_j^2}. \tag{51c}$$

Then, it is not difficult to see that

$$L(Y) \doteq 1 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}e^{-\frac{(\tilde{\mu}_i - \tilde{\mu}_j)^2}{2(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}. \tag{52}$$

where $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ (or $\tilde{\mu}_j$ and $\tilde{\sigma}_j^2$) denote the means and the variances in Gaussian mixture model mentioned in Equation (48).

Furthermore, in the light of Equations (47) and (52), we have the message importance transfer measure with the constrained variances σ_k^2 as follows

$$\begin{aligned} C(\mu_x, \sigma_x^2) &= \max_{P(X)}\{L(Y)\} - L(Z) \\ &\doteq \frac{1}{2\sqrt{\pi\sigma_z^2}} - \min_{P_X(x)}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{e^{-\frac{(\tilde{\mu}_i - \tilde{\mu}_j)^2}{2(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}}{\sqrt{2\pi(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}\right\}. \end{aligned} \tag{53a}$$

$$s.t. \quad \sum_{k=1}^{N}\sigma_k^2 = N\Theta, \tag{53b}$$

where the parameter Θ can be regarded as a constant which is controlled by the system parameters σ_x^2, μ_x and μ_k, as follows

$$\Theta = \frac{1}{N}\sum_{k=1}^{N}\sigma_k^2 = \sigma_x^2 + \mu_x^2 - \frac{1}{N}\sum_{k=1}^{N}\mu_k^2. \tag{54}$$

Moreover, the parameter σ_z^2 is a system constant and μ_k are regarded as constants, while the parameters σ_k^2 $(k = 1, 2, \ldots, N)$ can be adjusted flexibly. According to the Lagrange multiplier method, when $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_N^2 = \Theta$, we have

$$\min_{P_X(x)}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{e^{-\frac{(\tilde{\mu}_i - \tilde{\mu}_j)^2}{2(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}}{\sqrt{2\pi(\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2)}}\right\} = \frac{1}{2N^2\sqrt{\pi(\Theta + \sigma_z^2)}}\sum_{i=1}^{N}\sum_{j=1}^{N}e^{-\frac{(\mu_i - \mu_j)^2}{4(\Theta + \sigma_z^2)}}. \tag{55}$$

By substituting Equation (55) into Equation (53a), the proof of Corollary 1 is already completed. □

In order to investigate the continuous information transfer processing mentioned in Corollary 1, we do some simulations shown as Figures 2 and 3. In particular, Figure 2 shows that when the variable X following a Gaussian mixture model transfers to the variable Y, the message importance measures of X and Y become more absolutely close with N increasing (N denotes the number of Gaussian functions in the Gaussian mixture model). Besides, we also see that the differences of message importance measures between the variable X and Y are not significant in the case of large variances σ_k^2. In addition, from Figure 3, it is seen that the message importance transfer capacity is increasing with the increment of the number of Gaussian functions. Moreover, the larger variances σ_k^2 of the Gaussian mixture model are, the larger message importance transfer capacity we have.

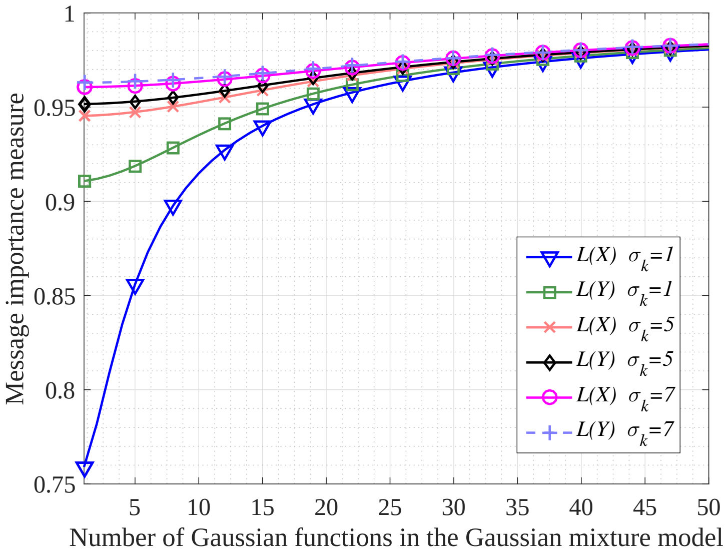


Figure 2. The comparison between the message importance measures for the original variable X and the final variable Y in an information transfer processing (where the variable X follows a Gaussian mixture model with all the variances of Gaussian functions as same as σ_k^2).

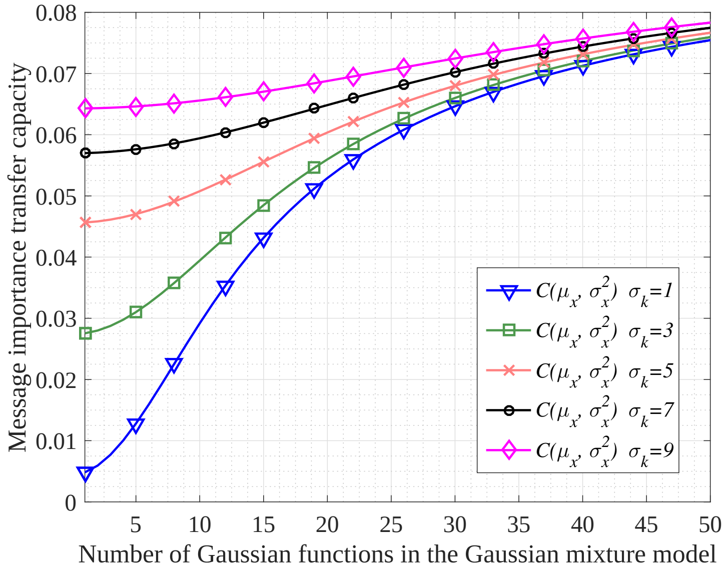


Figure 3. The performance of message importance transfer capacity in the Gaussian mixture model in which all the variances of Gaussian functions are the same as σ_k^2 ($\sigma_k = 1, 3, 5, 7, 9$).

Remark 4. *As for an additive disturbance system where the data source derive from a Gaussian mixture model, we can obtain the message importance transfer capacity, if there are all the same variances σ_k^2 for the Gaussian distribution components in the data source. In practice, when the power of signal source is controlled in a signal transfer processing, we can adjust signal distributions to achieve the optimal message importance transfer by using Corollary 1.*

4. Application in Mobile Edge Computing with the M/M/s/k Queue

As for mobile users, almost all of them have few computing resources and depend solely on cloud computing. This implies that the large distance between the cloud and the end devices is not suitable for the low delay requirement of the future applications. To cope with the issue, the MEC is proposed to improve cloud computing.

As far as the MEC is concerned, the edge servers are placed in the Base Stations (BSs) to reduce the delay, while context aware applications are close to the mobile users [34]. To characterize the MEC more specifically, a MEC model is constructed based on the queuing theory as follows.

In terms of a MEC system in Figure 4, it consists of many mobile users, an edge server, and a central cloud located far from the local devices. For each mobile user, a part of or all the service requests can be offloaded to the corresponding edge server when the communication is disturbed by other mobile users or environmental noise. If the upper bound of the service rate for the edge server is larger than the sum of mobile users' request rate, the offloaded requests will be coped with by the edge server. Otherwise, the overloaded requests will be offloaded to the central cloud for processing [35]. In these cases, the queue model on the edge server can be considered as the M/M/s/k queue, where the first M describes the request interarrival time of mobile users, the second M denotes the request service time in the edge server, and both of them follow exponential distribution; the parallel processing core number is s, which means each processing core can at most server one request simultaneously; the queuing buffer is k in the edge server. Note that we only consider a simple model on MEC to show the potential application of MITM. In fact, there may be some complicated cases in the MEC such as fault tolerance, failover, and the existence of overlay networks, etc.; we shall consider this in the near future.

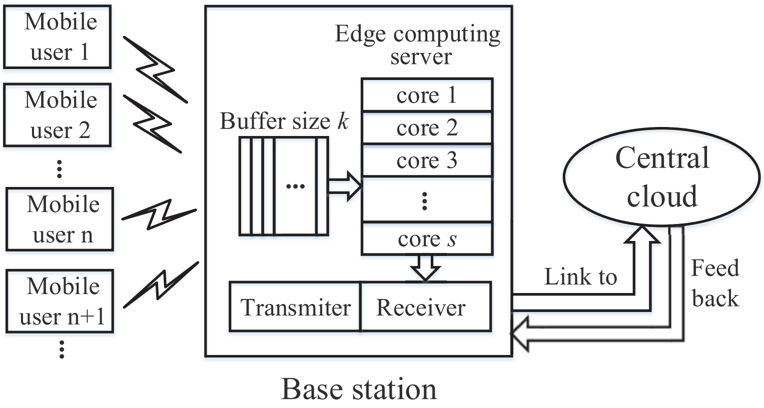


Figure 4. The queue model on the mobile edge computing system.

In fact, it is significant for the MEC system to use the finite buffer size (or caching size) to approximate the infinite one, which can be treated as a problem of queue length selection. To do this, we exploit the MITM and KL divergence to measure the effect of queue states variation on the MEC performance as follows.

4.1. MITM in the Queueing Model

As a measurement for the distance of the message importance, MITM characterizes the difference between two distributions. This can be applied to distinguish the state probability distributions in queue models. To give more general analysis, we discuss the relationship between the queue state stationary distributions in the M/M/s/k model. The queue state stationary probability of the model with arrival rate $\tilde{\lambda}$ and service rate $\tilde{\mu}$ can be described as

$$p_0 = \Big[\sum_{j=0}^{s-1} \frac{a^j}{j!} + \frac{a^s}{s!} \cdot \frac{1-\rho^{k+1}}{1-\rho}\Big]^{-1} \tag{56a}$$

$$p_j = \frac{a^j}{j!} p_0, \quad (0 < j < s) \tag{56b}$$

$$p_j = \frac{a^s}{s!} \rho^{j-s} p_0, \quad (s \le j \le s+k), \tag{56c}$$

where s is the number of servers, k is the size of buffer or cache, the traffic intensity $\rho = a\ /s < 1$ as well as $a = \tilde{\lambda}/\tilde{\mu}$.

Therefore, according to the definition 1, we can obtain the RMIM of the queue state stationary probability in the M/M/s/k model. Then, by use of Taylor series expansion, the approximate RMIM is given by

$$\begin{aligned} &\sum_{j=0}^{s+k} p_j e^{-p_j} \\ &= \sum_{j=0}^{s+k} p_j [1 - p_j + O(p_j^2)] \\ &\doteq 1 - \sum_{j=0}^{s-1} \frac{(\frac{a^j}{j!})^2}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} - \sum_{j=s}^{s+k} \frac{\varphi_2^2 \rho^{2j}}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} \\ &= 1 - p_0^2 \Big\{ \sum_{j=0}^{s-1} (\frac{a^j}{j!})^2 + \frac{\varphi_2^2 [1-\rho^{2(k+1)}]}{1-\rho^2} \Big\}, \end{aligned} \tag{57}$$

where the parameter φ_1 and φ_2 are

$$\varphi_1 = \sum_{j=0}^{s-1} \frac{a^j}{j!}, \tag{58a}$$

$$\varphi_2 = \frac{a^s}{s!}. \tag{58b}$$

Furthermore, referring to Equation (57), we can use the MITM to characterize the information difference for the queue model as follows.

Proposition 5. *In the M/M/s model, the MITM can be used to measure the information difference between two queue state probability stationary distributions $P_k = \{p_{k,0}, p_{k,1}, \ldots, p_{k,s+k}, 0, 0, \ldots, 0\}$ and $P_{k+1} = \{p_{k+1,0}, p_{k+1,1}, \ldots, p_{k+1,s+k+1}, 0, \ldots, 0\}$ which are with buffer size k and $k+1$ respectively, as follows*

$$\begin{aligned} D_I(P_{k+1}||P_k) &= \sum_{j=0}^{s+k+1} p_{k+1,j} e^{-p_{k+1,j}} - \sum_{j=0}^{s+k} p_j e^{-p_j} \\ &\doteq \Big(\frac{1}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} - \frac{1}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+2}}{1-\rho})^2}\Big) \sum_{j=0}^{s-1} (\frac{a^j}{j!})^2 \\ &+ \frac{\varphi_2^2}{1-\rho^2} \Big[\frac{1-\rho^{2(k+1)}}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} - \frac{1-\rho^{2(k+2)}}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+2}}{1-\rho})^2}\Big], \end{aligned} \tag{59}$$

where the constraint satisfies $|D_I(P_{k+1}||P_k)| \le \lambda \|P_{k+1} - P_k\|_1$ ($\lambda > 0$ is a constant), and φ_1 and φ_2 are mentioned in Equations (58a) and (58b).

Likewise, it is readily seen that the MITM between the queue state stationary probability distributions $P_\infty = \{p_{\infty,0}, p_{\infty,1}, \dots, p_{\infty,\infty}\}$ and $P_k = \{p_{k,0}, p_{k,1}, \dots, p_{k,s+k}, 0, 0, \dots, 0\}$ is given by

$$\begin{aligned} &D_I(P_\infty||P_k) \\ &= (\frac{1}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} - \frac{1}{(\varphi_1 + \frac{\varphi_2}{1-\rho})^2}) \sum_{j=0}^{s-1} (\frac{a^j}{j!})^2 + \frac{\varphi_2^2}{1-\rho^2} [\frac{1-\rho^{2(k+1)}}{(\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho})^2} - \frac{1}{(\varphi_1 + \frac{\varphi_2}{1-\rho})^2}]. \end{aligned} \tag{60}$$

In this case, the buffer selection problem in MEC can be formulated as

$$k_I^* = \min_k \{k; |D_I(P_\infty||P_k)| \leq \delta\}, \tag{61}$$

where $\delta > 0$ is the threshold of variation in former difference.

In particular, if there is only one server, the corresponding queue model is M/M/1/k, it is not difficult to obtain

$$D_{I(s=1)}(P_\infty||P_k) = \frac{2a}{1+a} - \frac{2a(1-a^{k+1})}{(1+a)(1-a^{k+2})}, \tag{62}$$

where $D_{I(s=1)}(P_\infty||P_k)$ denotes the MITM with the number of server $s = 1$. The corresponding optimal buffer size is given by

$$k^*_{I(s=1)} = \frac{\ln \frac{\delta(1+a)}{a(2+\delta-(2-\delta)a)}}{\ln a} - 1. \tag{63}$$

It is apparent that δ plays an important role in selecting the caching size when using finite size caching to imitate the infinite caching working mode.

4.2. KL Divergence in the Queue Model

As a common information measure, KL divergence is also considered to be applied to measuring the information distinction between the queue state stationary probability distributions with different buffer sizes in the queue models. In particular, for the M/M/s model, we have the following proposition.

Proposition 6. *In the M/M/s model, the KL divergence between the queue state distribution* $P_{k+1} = \{p_{k+1,0}, p_{k+1,1}, \dots, p_{k+1,s+k+1}, 0, \dots, 0\}$ *and* $P_k = \{p_{k,0}, p_{k,1}, \dots, p_{k,s+k}, 0, 0, \dots, 0\}$ *with buffer size* $k+1$ *and* k*, is derived as*

$$\begin{aligned} D(P_k||P_{k+1}) &= \sum_j p_{k,j} \log \frac{1}{p_{k+1,j}} - \sum_j p_{k,j} \log \frac{1}{p_{k,j}} \\ &= \sum_{j=0}^{s-1} \frac{a^j}{j!} p_{k,0} \log \frac{p_{k,0}}{p_{k+1,0}} + \sum_{j=s}^{s+k} \frac{a^s}{s!} \rho^{j-s} p_{k,0} \log \frac{p_{k,0}}{p_{k+1,0}} \\ &= \log \left\{ 1 + \frac{\varphi_2 \frac{(1-\rho)\rho^{k+1}}{1-\rho^{k+1}}}{\frac{\varphi_1(1-\rho)}{1-\rho^{k+1}} + \varphi_2} \right\}, \end{aligned} \tag{64}$$

where the parameters $p_{k,j}$*,* $p_{k+1,j}$*,* φ_1 *and* φ_2 *are the same as them in Proposition 5.*

Furthermore, it is not difficult to have the KL divergence between the distribution $P_\infty = \{p_{\infty,0}, p_{\infty,1}, \dots, p_{\infty,\infty}\}$ and $P_k = \{p_{k,0}, p_{k,1}, \dots, p_{k,s+k}, 0, 0, \dots, 0\}$ with buffer size ∞ and k, which is obtained as

$$\begin{aligned} D(P_k||P_\infty) &= \sum_j p_{k,j} \log \frac{1}{p_{\infty,j}} - \sum_j p_{k,j} \log \frac{1}{p_{k,j}} \\ &= \log \frac{p_{k,0}}{p_{\infty,0}} \\ &= \log \frac{\varphi_1 + \varphi_2 \frac{1}{1-\rho}}{\varphi_1 + \varphi_2 \frac{1-\rho^{k+1}}{1-\rho}}. \end{aligned} \tag{65}$$

Similar to Equation (61), it is rational for us to use KL divergence as measurement to select the buffer size. The corresponding optimal buffer size can be described as

$$K^*_{KL} = \min_k \{k; |D(P_k||P_\infty)| \le \delta\}, \tag{66}$$

where $\delta > 0$ is a small enough parameter and it can adjust the information transfer gap between the queue state stationary distributions P_∞ and P_k which are with buffer size ∞ and k respectively. Then, we have

$$k^*_{KL} = \frac{\ln\left(1 - \frac{(1-\rho)\varphi_1 + \varphi_2 - 2^\delta \varphi_1 (1-\rho)}{2^\delta \varphi_2}\right)}{\ln \rho} - 1, \tag{67}$$

where k is the buffer size or queue length, φ_1 and φ_2 are mentioned in Equations (58a) and (58b). What is more, as for the M/M/1/k model, the optimal buffer size is simplified as follow

$$k^*_{KL} = \frac{\ln(1 - \frac{1}{2^\delta})}{\ln a} - 2. \tag{68}$$

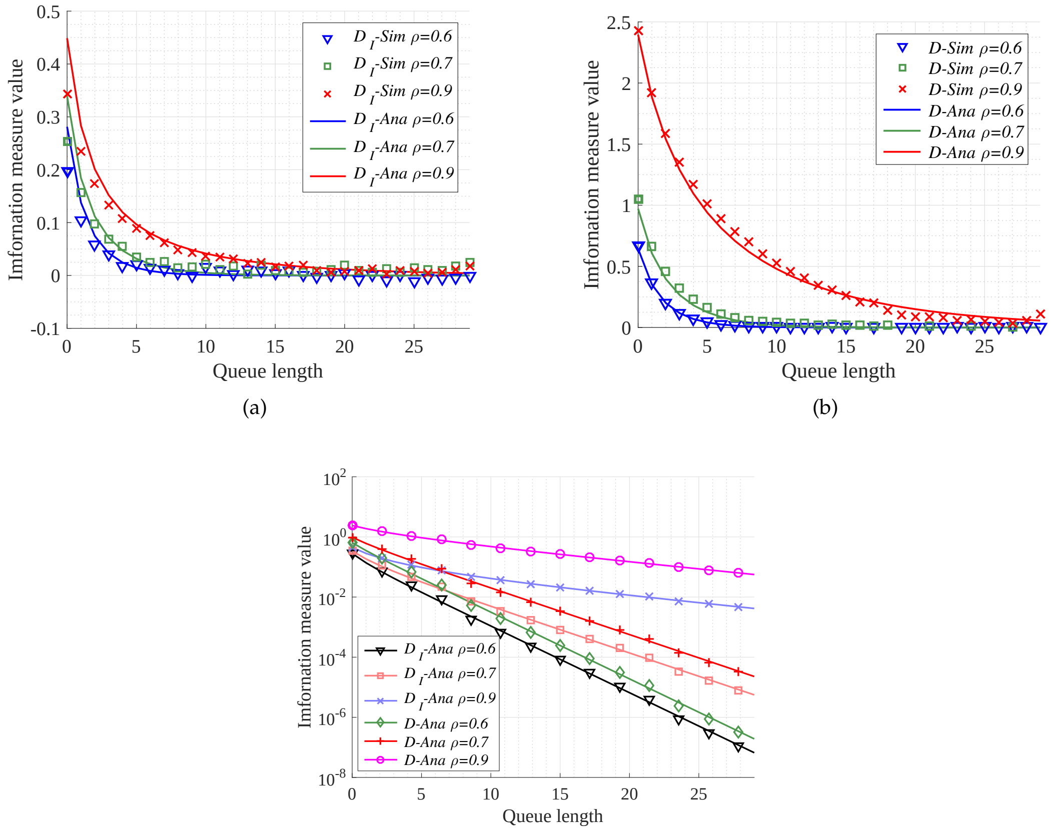
Therefore, by using the information measures such as MITM and KL divergence, it may provide an effective method to select the caching size, which can exploit the resources of MEC more reasonably.

4.3. Numerical Validation

To validate our derived results in theory, we take some event simulation experiments of the queue model by use of Matlab. By setting the arrival rate $\tilde{\lambda}$ and service rate $\tilde{\mu}$ of queue model, the process of arrival and departure for each event is simulated during a fixed period. We will elaborate on specific parameters of the queue model in the following context. In the figures of results, the legends D_I*-Sim*, D_I*-Ana* and *D-Sim*, *D-Ana* are used to denote the simulation results and the analytical results for MITM and KL divergence, respectively.

4.3.1. Effect of the Traffic Intensity

We now exploit M/M/s/k model to investigate performance of the MITM and KL divergence in the case of different traffic intensity. In the simulations, the queue length, namely the buffer size, is set to change from 0 to 30, the number of servers satisfies $s = 1$, and the traffic intensity is set as $\rho = 0.6, 0.7, 0.9$. Then, we can compare the simulation results with the theoretical ones for the MITM and KL divergence. From Figure 5, it is seen that the analytical results mentioned in Equations (59) and (60) can validate the simulation results. In particular, Figure 5a,b shows that analytical results of MITM and KL divergence can absolutely fit the simulation experiments in the M/M/s/k models with different traffic intensity. What is more, from Figure 5c, we can see that in the same queue model the convergence for MITM is faster than that for KL divergence. That is, the MITM offers a reasonably lower bound for queue length selection with respect to MEC. Besides, the less traffic intensity we have, the more caching size resources can be saved.



(a)

(b)

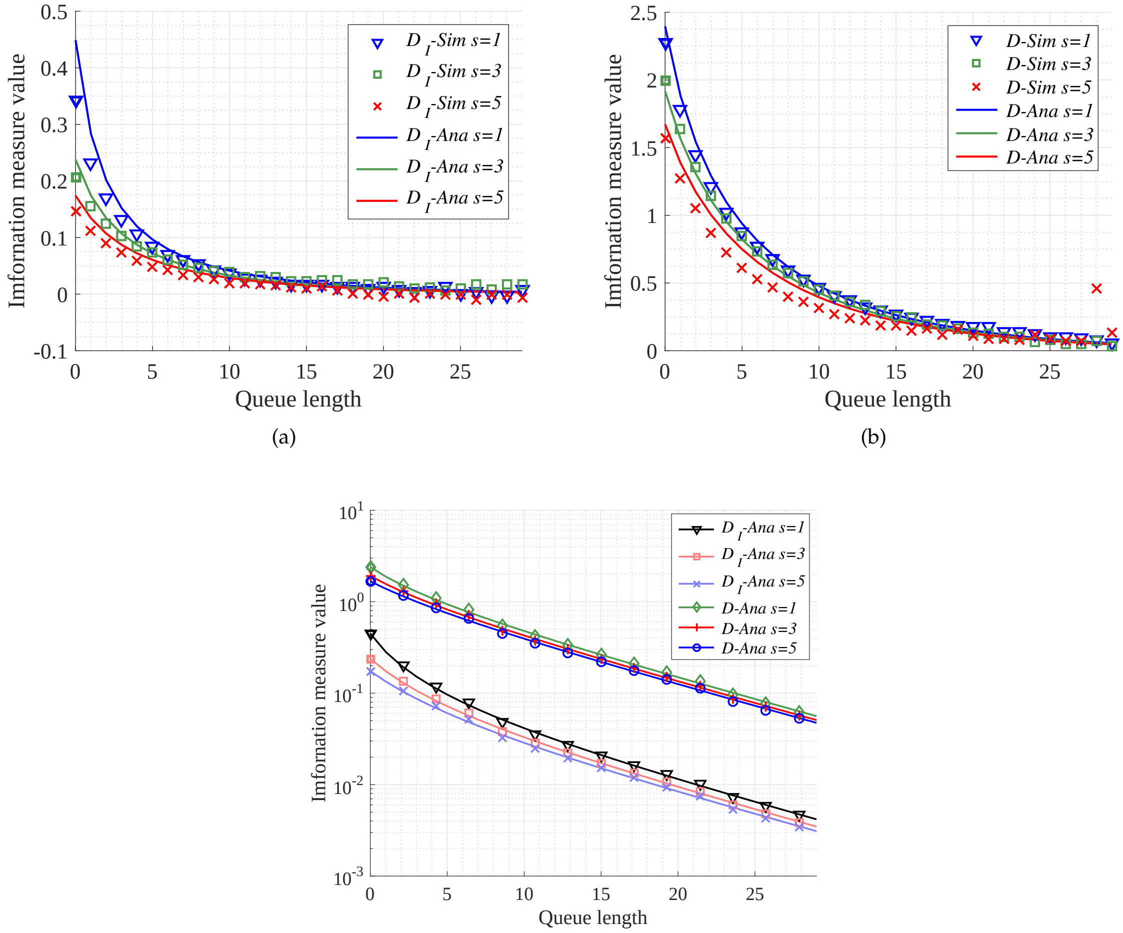
(c)

Figure 5. The performance of different information measures versus queue length. The queue models are with the same number of server $s = 1$ and different traffic intensity ρ (ρ = 0.6, 0.7, 0.9). (**a**) The performance of message importance transfer measure (MITM) mentioned in Equation (60) in the case of traffic intensity $\rho = 0.6, 0.7, 0.9$; (**b**) The performance of KL divergence mentioned in Equation (65) in the case of traffic intensity $\rho = 0.6, 0.7, 0.9$; (**c**) The analysis results of different information measures between the queue length k and ∞.

4.3.2. Effect of the Number of Servers

With regard to effects of number of servers on the MITM and KL divergence, we do the simulation experiments with M/M/s model by setting the number of servers as $s = 1, 3, 5$. What is more, we set the queue length as $k = 0, 1, 2, \ldots 30$, and the traffic intensity always as $\rho = 0.9$. Then, we gain the comparison between the simulation results and the theoretical ones.

Figure 6a,b show that it is almost available for analytical results to fit simulation results. From Figure 6c, similar to Figure 5c, we can also use the MITM to gain a lower bound for queue length selection than KL divergence. Moreover, keeping other conditions the same, a larger number of servers can make MITM and KL divergence converge faster. In other words, there is a trade-off between the number of servers and caching size.



(a)

(b)

(c)

Figure 6. The performance of different information measures versus queue length. The queue models are with the same traffic intensity $\rho = 0.9$ and different number of servers (s = 1, 3, 5). (**a**) The performance of MITM mentioned in Equation (60) in the case of the number of servers $s = 1, 3, 5$; (**b**) The performance of KL divergence mentioned in Equation (65) in the case of the number of servers $s = 1, 3, 5$; (**c**) The analysis results of different information measures between the queue length k and ∞.

4.3.3. Performance Results for Different Arrival Events Distributions

Now we discuss the performance results in the cases of different distributions of events' arrivals which is listed in Table 1. It is apparent that average interarrival time is maintained as the same, namely $1/\tilde{\lambda}_0$. As well, we make sure that the number of server and traffic intensity are $s = 1$ and $\rho = 0.9$ in all cases, respectively. Then, we make simulations in the three cases to compare the testing results with the analytical results.

Table 1. The interarrival time distributions of events' arrivals.

Type of Distribution	**Exponential Distribution**	**Uniform Distribution**	**Normal Distribution**
$P(X)$	$X \sim E(\tilde{\lambda}_0)$	$X \sim U(0, 2/\tilde{\lambda}_0)$	$X \sim N(\frac{1}{\tilde{\lambda}_0}, \frac{1}{\tilde{\lambda}_0^2})$

As for Figure 7, it is illustrated that the convergence of MITM is faster than that of KL divergence, which indicates that MITM may provide a reasonable lower bound to select the caching size for MEC.

In addition, we can see that the Poisson distribution (namely, events' arrivals follow exponential distribution) corresponds to the worst case for the arrival process among the three discussed cases with respect to the convergence of both MITM and KL divergence.

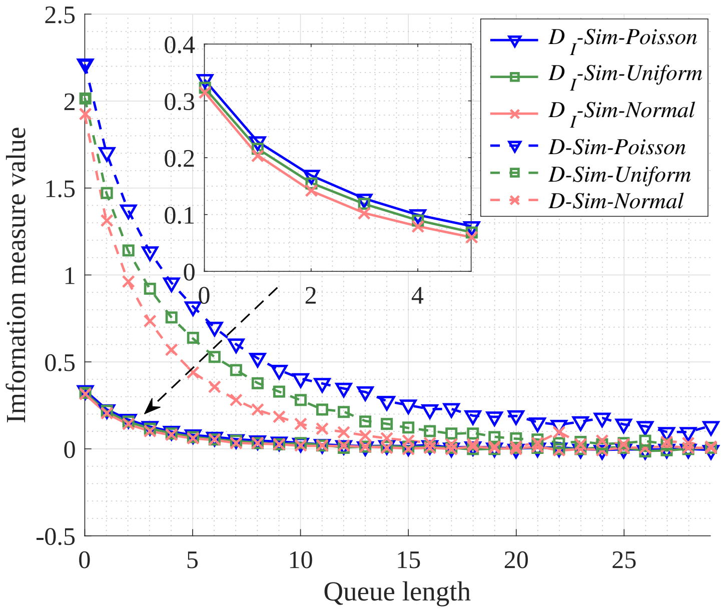


Figure 7. The performance of different information measures between the queue length k and ∞ for the queue models with the same number of server $s = 1$, the same traffic intensity $\rho = 0.9$, and the different arrival events' distributions.

5. Conclusions

In this paper, the information transfer problem was investigated from the perspective of information theory and big data analytics. An information measure, i.e., MITM, was proposed to characterize the information distance between two distributions, similar to KL divergence and Renyi divergence. Actually, the information measure plays a vital role in focusing on the message importance hidden in small probability events of big data. Therefore, it is applicable for the information measure to characterize information transfer process in big data. We have investigated the variation of message importance in the information transfer process by using MITM. Furthermore, we proposed the message importance transfer capacity based on the MITM so that an upper bound can be presented for the information transfer process with disturbance. In addition, we applied the information transfer measure to select the caching size in MEC. As the next step of research, we shall carry out real data experiments to test some of the most complicated cases of MEC and make use of the information transfer measures to investigate some related algorithms as well as to discuss the effect of window length on the whole system performance in big data analytics.

Author Contributions: R.S., S.L. and P.F. conceived and designed the methodology; R.S. and P.F. the mathematical analysis and the practical simulations; R.S., S.L. and P.F. discussed the results; R.S. and P.F. wrote the paper; All authors have read and approved the final manuscript.

Acknowledgments: The authors indeed thank a lot for the support of the China Major State Basic Research Development Program (973 Program) No. 2012CB316100(2), as well as the National Natural Science Foundation of China (NSFC) No. 61771283 and No. 61621091.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper. Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-whitepaper-c11-520862.html (accessed on 21 May 2017).
2. TechRepublic. Cloud Traffic to Jump 262% by 2020. Available online: http://www.techrepublic.com/article/cloud-traffic-to-jump-262-by-2020-according-to-cisco-global-cloud-index (accessed on 21 May 2017).
3. Ju, B.; Zhang, H.; Liu, Y.; Liu, F.; Lu, S.; Dai, Z. A feature extraction method using improved multi-scale entropy for rolling bearing fault diagnosis. *Entropy* **2018**, *20*, 212. [CrossRef]
4. Wei, H.; Chen, L.; Guo, L. KL divergence-based fuzzy cluster ensemble for image segmentation. *Entropy* **2018**, *20*, 273. [CrossRef]
5. Rehman, S.; Tu, S.; Rehman, O.; Huang, Y.; Magurawalage, C.M.S.; Chang, C.C. Optimization of CNN through novel training strategy for visual classification problems. *Entropy* **2018**, *20*, 273. [CrossRef]
6. Chen, X.W.; Lin, X.T. Big data deep learning: challenges and perspectives. *IEEE Access* **2014**, *2*, 514–525. [CrossRef]
7. Ramaswamy, S.; Rastogi, R.; Shim, K. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Rec.* **2000**, *29*, 427–438. [CrossRef]
8. Lee, W.; Stolfo, S.J. Data Mining Approaches for Intrusion Detection. In Proceedings of the Usenix security, San Antonio, TX, USA, 26–29 January 1998; pp. 291–300.
9. Julisch, K.; Dacier, M. Mining intrusion detection alarms for actionable knowledge. In Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 366–375.
10. Wang, S. A comprehensive survey of data mining-based accounting-fraud detection research. In Proceedings of the IEEE Intelligent Computation Technology and Automation (ICICTA), Madurai, India, 11–12 May 2010; pp. 50–53.
11. Zieba, A. Counterterrorism systems of spain and poland: comparative studies. *Przeglad Politol.* **2015**, *3*, 65–78. [CrossRef]
12. Phua, C.; Lee, V.; Smith, K.; Gayler, R. A comprehensive survey of data mining-based fraud detection research. *arXiv* **2010**, arXiv:1009.6119. [CrossRef]
13. Ando, S. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In Proceedings of the 7th International Conference on Data Mining, Omaha, NE, USA, 28–31 October 2007; pp. 13–22.
14. Ando, S.; Suzuki, E. An information theoretic approach to detection of minority subsets in database. In Proceedings of the 6th International Conference on Data Mining, Hong Kong, China, 18–22 December 2006; pp. 11–20.
15. He, J.; Liu, Y.; Lawrence, R. Graph-based rare category detection. In Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 418–425.
16. Zhou, D.; Wang, K.; Cao, N.; He, J. Rare category detection on time-evolving graphs. In Proceedings of the 15th IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015; pp. 1550–4786.
17. Fan, P.; Dong, Y.; Lu, J.; Liu, S. Message importance measure and its application to minority subset detection in big data. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
18. Renyi, A. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1961; Volume 1, pp. 547–561.
19. Carter, K.M.; Raich, R.; Hero, A.O. On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Process.* **2010**, *58*, 650–663. [CrossRef]
20. Anderson, A.; Haas, H. Kullback-Leibler Divergence (KLD) based anomaly detection and monotonic sequence analysis. In Proceedings of the IEEE Vehicular Technology Conference (VTC Fall), San Francisco, CA, USA, 5–8 September 2011; pp. 1–6.
21. Chai, B.; Walther, D.; Beck, D. Exploring functional connectivities of the human brain using multivariate information analysis. In Proceedings of the IEEE AAnnual Conference on Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 9–12 December 2009; pp. 1–6.

22. She, R.; Liu, S.; Dong, Y.; Fan, P. Focusing on a probability element: parameter selection of message importance measure in big data. In Proceedings of the IEEE International Conference on Communications (ICC), Paris, France, 20–26 May 2017; pp. 1–6.
23. Liu, S.; She, R.; Fan, P.; Letaief, K.B. Non-parametric Message Important Measure: Storage Code Design and Transmission Planning for Big Data. Available online: https://arxiv.org/abs/1709.10280 (accessed on 29 September 2017).
24. She, R.; Liu, S.; Fan, P. Amplifying inter-message distance: On information divergence measures in big data. *IEEE Trans. Signal Process.* **2017**, *58*, 24105–24119. [CrossRef]
25. Massey, J.L. Causality, feedback and directed information. In Proceedings of the International Symposium on Information Theory and its Applications, Waikiki, HI, USA, 27–30 November 1990; pp. 1–6.
26. Kramer, G. Directed Information for Channels With Feedback. Ph.D. Thesis, Swiss Federal Institute of Technology Zurich, Zurich, Switzerland, 1998.
27. Zhao, L.; Kim, Y.H.; Permuter, H.H.; Weissman, T. Universal estimation of directed information. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Austin, TX, USA, 13–18 June 2010; pp. 230–234.
28. Charalambous, C.D.; Kourtellaris, C.K.; Tzortzis, I. Information transfer of control strategies: Dualities of stochastic optimal control theory and feedback capacity of information theory. *IEEE Trans. Autom. Control* **2017**, *62*, 5010–5025. [CrossRef]
29. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. [CrossRef] [PubMed]
30. Sinha, S.; Vaidya, U. Causality preserving information transfer measure for control dynamical system. In Proceedings of the IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, USA, 12–14 December 2016; pp. 7329–7334.
31. Sinha, S.; Vaidya, U. Formalism for information transfer in dynamical network. In Proceedings of the IEEE 54th Annual Conference on Decision and Control (CDC), Osaka, Japan, 15–18 December 2015; pp. 5731–5736.
32. Liang, X.S.; Kleeman, R. Information transfer between dynamical system components. *Phys. Rev. Lett.* **2000**, *95*, 1–4. [CrossRef] [PubMed]
33. Huang, S.; Makur, A.; Zheng, L.; Wornell, G.W. An information-theoretic approach to universal feature selection in high-dimensional inference. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 1336–1340.
34. Ndikumana, A.; Ullah, S.; LeAnh, T.; Tran, N.H.; Hong, C.S. Collaborative cache allocation and computation offloading in mobile edge computing. In Proceedings of the Asia-Pacific Network Operations and Management Symposium (APNOMS), Seoul, Korea, 27–29 September 2017; pp. 366–369.
35. Liu, L.; Chang, Z.; Guo, X.; Ristaniemi, T. Multi-objective optimization for computation offloading in mobile-edge computing. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 3–6 July 2017; pp. 832–837.