Integer-Granularity Locality-Sensitive Bloom Filter

Jiangbo Qian, *Member, IEEE*, Qiang Zhu, *Senior Member, IEEE*, and Huahui Chen

***Abstract*— Numerous network applications require a fast and efficient way to check whether a given (query) object is close to at least one data object in a particular set to reduce unnecessary transmissions. Some variants of the Bloom filter, such as the multi-granularity locality-sensitive Bloom filter (MLBF), have been suggested to meet this requirement. Nevertheless, the MLBF was designed to only filter (query) objects with multiple logarithmic distance granularities. In this letter, we propose a novel Bloom filter structure, called the integer-granularity locality-sensitive Bloom filter (ILBF), which can filter objects with multiple integer distance granularities. Theoretical analyses for the property of the ILBF are presented. Experiments show that the theoretical estimates are quite accurate and the ILBF structure is efficient and effective.**

***Index Terms*— Bloom filter, locality-sensitive hashing, false positive/negative rates.**

I. Introduction

NUMEROUS network applications require a fast and efficient way to check whether a given (query) object is close to at least one data object in a particular set to reduce unnecessary transmissions. For example, let us consider an application scenario in which a music Web site needs to check if each music file submitted by a user contains any copyrighted contents. It is observed that most of the submissions have no copyright violation. Instead of sending each submitted music file to a remote Web service for a detailed copyright analysis, the Web site can perform a preliminary filtering by utilizing an on-site efficient compact filter (e.g., a Bloom filter) which is created based on a copyright music data set. In this way, most legal music submissions can be quickly filtered out without being sent out for an expensive remote examination. The comparison distance values for the filtering may vary, depending on the regions and applicable rules for the concerned users and Web site. This mechanism will significantly improve the overall performance of a network system in such applications. Other application examples include animal conservation, terrorist detection, and medical image processing.

The DSBF (distance-sensitive Bloom filter) [1] was the first filter towards the above objective. It is an integrated method of the locality-sensitive hashing (LSH) [2] and the Bloom filter (BF) [3], [4] for filtering data in the Hamming space. The DSBF achieves improvement in both time and space, comparing to the method that performs comparisons of the given filter/query object against the entire given data object set. Based on the LSH and the BF structure, the LSBF [5] employs an additional verification BF to further decrease the space cost and the false positive rate.

One limitation for the above two techniques is that they can filter only for one predetermined distance value. This value is a key parameter that indicates how close a filter/query object is to at least one of the (data) objects in the given set. Once a distance value is determined, it cannot be changed unless the BF structure is rebuilt. Unfortunately, the space cost for rebuilding is very high since it requires keeping the given set of data objects. Furthermore, the computing cost for rebuilding may not be acceptable either especially for a low end or overwhelmingly busy computing device like a smart field monitor/sensor in an animal conservation application or a busy Web server machine used for the aforementioned copyright protection application. On the other hand, the approach for maintaining multiple BF structures with different distance values may suffer drawbacks such as space overhead and inconsistent issue. To solve the problem, Qian et al. proposed the MLBF (multi-granularity locality-sensitive Bloom filter) and the MLBF* (a more space-effective variant of MLBF) [6] to filter objects with respect to multiple distances without having to rebuild the filter structure for each distance value. However, although the MLBF/MLBF* is a big improvement over the existing one-distance solutions, it was designed for only filtering objects with multiple logarithmic granularities (i.e., 1, 2, 4, 8, 16,...,2^b ($b = 0, 1, 2,...$)). For a given arbitrary distance value x (e.g., 9), a conservative coarser logarithmic value (e.g., 16) has to be used to approximate the request. As a result, many unwanted objects with respect to a given distance value x may survive the filtering. A filter with a finer granularity such as integers (1, 2, 3, ..., 8, 9, 10, ...) would provide a more effective filtering to better meet the user's requirement.

To overcome the above problem, in this letter, we propose a new Bloom filter technique, called the integer-granularity locality-sensitive Bloom filter (ILBF). The main contributions of this letter are the following: (1) a novel ILBF structure which can filter objects with multiple integer granularities for their closeness evaluation is proposed; (2) theoretical analyses for choosing proper LSH functions and estimating false positive and negative rates are presented; (3) experiments show that the theoretical estimates are quite accurate and the ILBF structure is efficient and effective.

II. ILBF Structure

Unlike a simple extension of the MLBF, the ILBF adopts a totally different approach. It utilizes the following function.

Definition 1 (Alignable LSH-Mod Function $\mathcal{H}$): A function in the alignable LSH-mod family $\mathcal{H}$ is defined as $h(\mathbf{x}) = \lfloor \mathbf{a} \cdot \mathbf{x}/w \rfloor \% m$, where $\mathbf{x}$ is the vector representation of an object in R^d, $\mathbf{a}$ is a d-dimensional vector whose elements

Manuscript received July 12, 2016; accepted August 3, 2016. Date of publication August 16, 2016; date of current version November 9, 2016. This work was supported by China NSF Grant No. 61472194 and 61572266, Ningbo NSF Grant No. 2014A610023 as well as programs sponsored by K. C. Wong Magna Fund in Ningbo University. The associate editor coordinating the review of this letter and approving it for publication was Z. Ding.

J. Qian and H. Chen are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China (e-mail: qianjiangbo@nbu.edu.cn; chenhuahui@nbu.edu.cn).

Q. Zhu is with the Department of Computer and Information Science, University of Michigan, Dearborn, MI 48128 USA (e-mail: qzhu@umich.edu).

Digital Object Identifier 10.1109/LCOMM.2016.2600670

1558-2558 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

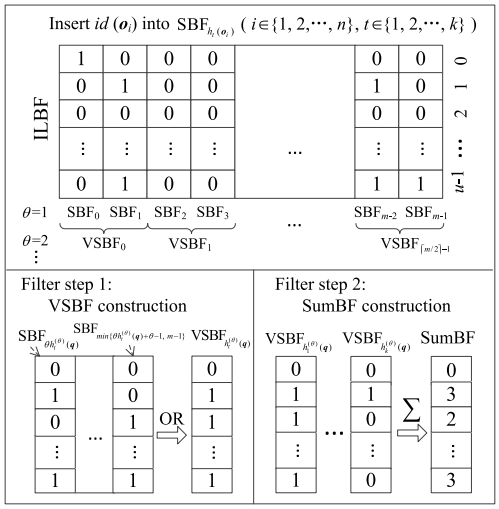

Fig. 1. The ILBF structure.

are chosen independently from a p-stable distribution, w is a user-specified constant for the filter/query distance, and m is a modulo parameter.

The ILBF structure illustrated in Fig.1 is composed of m standard Bloom filters (SBF) with each having u locations and k' hash functions. The procedure for inserting n data objects from a given set via the ILBF works as follows: (1) assigning each data object with a unique ID; (2) using k alignable LSH-mod functions to hash each object into [0, m) to choose k SBFs first; (3) then hashing the ID of each object into its chosen k SBFs, that is, setting k' bits to '1' in each of those k SBFs.

Let θ (=1,2,3,...,) be a granularity parameter related to different filter distance values w, $2w$, $3w$.... From Definition 1, we can see that, if parameter w is changed to θw, we essentially construct a (virtual) coaser granularity ILBF (using θw) from the finest granularity ILBF (using w) by combining every θ SBFs (i.e., SBF$_{z*\theta}$, SBF$_{z*\theta+1}$, ..., SBF$_{min\{(z*\theta+\theta-1),(m-1)\}}$, where $z = 0, 1, 2, ..., \lceil \frac{m}{\theta} \rceil - 1$) into a virtual SBF (i.e., VSBF$_z$). Let $h^{(\theta)}(\mathbf{x})$ denote the alignable LSH-mod function obtained by replacing w by θw in $h(\mathbf{x})$ from Definition 1. If a filter/query object $\mathbf{q}$ is close to a data object $\mathbf{o}$ (with respect to distance parameter θw), they have a high chance to be hashed into the same VSBF when using each $h^{(\theta)}(\mathbf{x})$. In other words, the k SBFs chosen to store the data object $\mathbf{o}$ by using $h(\mathbf{o})'s$ are likely contained in the relevant VSBFs by using the corresponding k $h^{(\theta)}(\mathbf{q})'s$. Based on this observation, we can apply the following criterion to answer a query: query object $\mathbf{q}$ has a close data object if we can find a data object $\mathbf{o}$ that has a significant number υ of SBFs (among the k SBFs chosen for $\mathbf{o}$) that are contained in the corresponding k VSBFs chosen for $\mathbf{q}$. Since the ID of data object $\mathbf{o}$ is stored these υ SBFs, if we bitwise sum up VSBFs to constitute a SumBF, at least k' locations of the SumBF are not less than the value υ.

Hence, the filter/query procedure works as follows: (1) using k alignable LSH-mod functions $h^{(\theta)}(\mathbf{x})$'s (corresponding to k original $h(\mathbf{x})$'s for choosing SBFs) to hash the query object to identify the relevant k VSBFs; (2) using a bitwise OR operation to combine the relevant θ SBFs and form each VSBF; (3) using a bitwise summation operation to sum up the k VSBFs to form a summary filter SumBF; (4) returning a positive answer if at least k' locations of the SumBF are not less than a threshold value υ.

III. Theoretical Analysis

Theorem 1: The collision probability of two objects $\mathbf{o}_i$ and $\mathbf{q}$ with distance $c_i = \|\mathbf{o}_i - \mathbf{q}\|$ under an LSH-mod function $h(\mathbf{x}) = \lfloor \mathbf{a} \cdot \mathbf{x}/w \rfloor \%m$ is $(\int_0^w ((\frac{1}{c_i} f(\frac{t}{c_i}))\%(wm))(1 - \frac{t}{w})dt + \int_{wm-w}^{wm} ((\frac{1}{c_i} f(\frac{t}{c_i}))\%(wm))(\frac{t-wm+w}{w})dt)/\int_0^{wm} ((\frac{1}{c_i} f(\frac{t}{c_i}))\%(wm))dt$, where $f(t)$ denotes the probability density function of the absolute value of the standard normal distribution, i.e., $f(t) = \sqrt{\frac{2}{\pi}} e^{\frac{-t^2}{2}}$, $t \geq 0$.

Sketch of Proof: Because the result of the module operation is positive, we have $t = |(\mathbf{a} \cdot \mathbf{o}_i)\%(wm) - (\mathbf{a} \cdot \mathbf{q})\%(wm)| = (\mathbf{a} \cdot \mathbf{o}_i - \mathbf{a} \cdot \mathbf{q})\%(wm)$. As a random vector $\mathbf{a}$ whose component values are drawn from the standard normal distribution, t is distributed as $(c_i|X|)\%(wm)$ from the Stable Distribution Theory, where $|X|$ is a random variable drawn from the standard normal distribution, i.e., $|X| \sim N(0, 1)$. From [6], the probability of $(c_i|X|)\%(wm) = t$ is $((1/c_i) f(t/c_i))\%(wm)$, where $f(t)$ is the probability density function of $|X|$. As the module operation, there are two cases for $\mathbf{o}_i$ and $\mathbf{q}$ falling into the same bucket: (1) when $t \in (0, w)$, the probability is $(w - t)/w = 1 - t/w$; (2) when $t \in (wm - w, wm)$, the probability is $(t - (wm - w))/w$. Hence, Theorem 1 holds. □

Intuitively, if two objects are close to each other and the product of wm is large, the module operation can be omitted. If two objects are far away from each other and the product of wm is small, the module operation can be viewed as a random mapping operation in [0, m). Therefore, we have:

Observation 1: The collision probability of two objects $\mathbf{o}_i$ and $\mathbf{q}$ with distance $c_i = \|\mathbf{o}_i - \mathbf{q}\|$ under an LSH-mod function $h(\mathbf{x}) = \lfloor \mathbf{a} \cdot \mathbf{x}/w \rfloor \%m$ can be estimated as: (1)$\int_0^w \frac{1}{c_i} f\left(\frac{t}{c_i}\right)\left(1 - \frac{t}{w}\right) dt$, if $\int_0^w \frac{1}{c_i} f\left(\frac{t}{c_i}\right)\left(1 - \frac{t}{w}\right) dt \geq \frac{1}{m}$; or (2)$\frac{1}{m}$, if $\int_0^w \frac{1}{c_i} f\left(\frac{t}{c_i}\right)\left(1 - \frac{t}{w}\right) dt < \frac{1}{m}$.

Theorem 1 and Observation 1 will be verified by experiments in Section IV.

Theorem 2: For a filter object $\mathbf{q}$ and k LSH-mod functions, assume that the query result is positive if the counts in at least k' locations of the SumBF are not smaller than parameter value υ. If, among all the $\mathbf{o}_i{}'$ s, only one object (assume $\mathbf{o}_i$) is close enough to the given filter object $\mathbf{q}$ with collision probability of p_i, the false negative rate can be estimated as FNR=$\sum_{j=0}^{\upsilon-1} \binom{k}{j} p_i^j (1 - p_i)^{k-j}$.

Sketch of Proof: For a filter object $\mathbf{q}$ and k LSH-mod functions, if $\mathbf{q}$ collides with an object for k times, the counts in at least k' locations of the SumBF are not smaller than υ. Thus, a false negative occurs when the number of times for collision is smaller than υ. The probability of collision for 0 time is $\binom{k}{0} p_i^0 (1 - p_i)^k$ and the probability of collision for 1 time is $\binom{k}{1} p_i^1 (1 - p_i)^{k-1}$, and so on. Therefore, Theorem 2 holds. □

Theorem 3: The false positive rate of the ILBF can be estimated as: FPR=1-$\sum_{j=0}^{k'-1} \binom{u}{j} p_{(\upsilon)}^j (1 - p_{(\upsilon)})^{u-j}$, where $p_{(\upsilon)} = 1 - \sum_{j=0}^{\upsilon-1} \binom{k^2 nk'/m}{j} \left(\frac{1}{u}\right)^j (1 - \frac{1}{u})^{k^2 nk'/m - j}$.

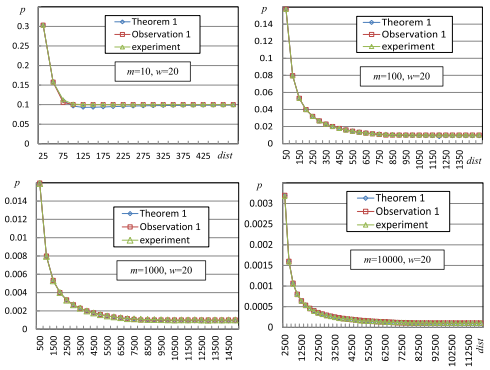

Fig. 2. Collision probability of the LSH-mod function.

Sketch of Proof: If two objects are far away from each other, from Observation 1, they will be uniformly distributed in SBFs with indexes in range $[0,\ m-1]$. Thus, from the mathematical expectation, there are kn/m objects in one SBF and then k^2n/m objects in k SBFs. In our method, the '1' or '0' bits at the same number of locations in those k SBFs are summed up to construct the SumBF. If the locations (in the same SBF) mapped from objects are not identical, the SumBF could be regarded as an approximate counting BF with u locations and k^2n/m objects because the maximum count at one location of an SBF is $'1'$. The probability of one counter in the SumBF being equal to or greater than parameter υ is: $p_{(\upsilon)} = 1 - \sum_{j=0}^{\upsilon-1} \binom{k^2nk'/m}{j} \left(\frac{1}{u}\right)^j \left(1-\frac{1}{u}\right)^{k^2nk'/m-j}$. From the algorithm, if counters in the SumBF are equal to or greater than parameter υ, a false positive occurs. Therefore, the false positive rate is FPR=1-$\sum_{j=0}^{k'-1} \binom{u}{j} p_{(\upsilon)}^j (1-p_{(\upsilon)})^{u-j}$. □

IV. Experimental Results

We conducted extensive experiments on an i7/8G computer to evaluate the accuracy of the ILBF structure using synthetic data. All experimental results were measured from a large number of repeated executions (10^6) to achieve accurate mathematical expectations. The synthetic data were randomly generated with 20 dimensions that follow the uniform distribution over [1, 1000]. For the theoretical estimation, we applied the Simpson's rule to calculate the integrals in the estimation models.

To verify Theorem 1 and Observation 1, we generate synthetic object pairs $(\mathbf{o}_1, \mathbf{o}_2)$ with different distances and then using randomly generated $h(\mathbf{x})'s$ to investigate the relevant collision probabilities. The results are summarized in Fig. 2, which demonstrate that Theorem 1 and Observation 1 are quite accurate. With the increasing of the distance between two objects, the collision probability decreases until it reaches the minimum collision probability of $1/m$.

To compare the false negative rates from the experiments and the theoretical estimation, we randomly generated filter objects close to one of the objects in the ILBF as test objects. The difference in each dimension of two objects was set to 0.1. Fig. 3 shows the results for different granularities $\theta = 1, 2, 3$ and 4. From the figure, we can see that the two curves from theoretical estimation and the experiments for the same parameters match very well. From the LSH-mod function, we can see the collision probability is positively correlated with θ. Therefore, the coarser the granularity (i.e., bigger θ), the lower

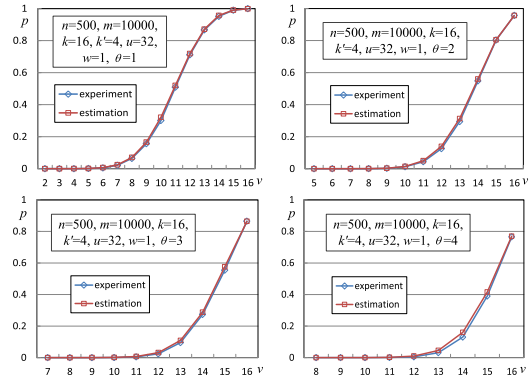

Fig. 3. FNRs of ILBF for different granularities.

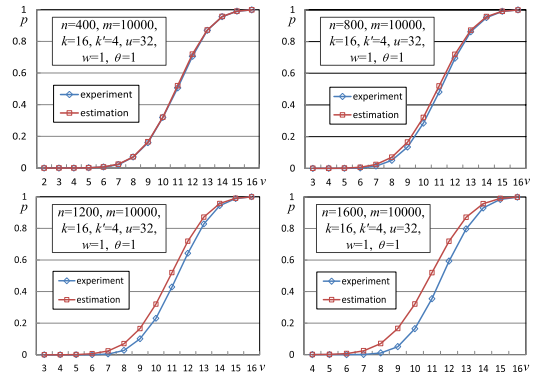

Fig. 4. FNRs of ILBF for different numbers of objects.

the false negative rate. For the same granularity, if υ is small, the false negative rates are low. Note that υ is a counter parameter for collision times. However, a small υ will also make the false positive rates high as shown by the following experiments. Hence, we should choose a proper parameter υ to make both the false positive rates and the false negative rate to be low.

Fig. 4 shows the results for different numbers of data objects in the ILBF. The two curves from the theoretical estimation and experiments exhibit a similar pattern. Similar to the standard BF, the FPR of the ILBF increases with the number of data objects. The increase of FPR will decrease the FNR. As the estimation from Theorem 2 does not consider the effect of the FPR, the estimation is larger than the observed results. We can choose a smaller parameter υ to control the false negative rate to a low level.

To compare the false positive rates from experiments and theoretical estimation, different numbers of synthetic data objects were stored in the ILBF, and additional synthetic filter objects which are not close to any of the objects in the ILBF were used as test objects. Fig. 5 shows the results. The two curves from the theoretical estimation and experiments exhibit a similar pattern. It can be seen that, the FPR is positively correlated with the number of objects in an ILBF. A larger parameter υ may be chosen to control the false positive rate to a low level.

To compare the false positive rates from experiments and theoretical estimation for granularities 1, 2, 3 and 4, respectively, 500 synthetic data objects were stored in the ILBF, and additional 10^5 synthetic filter objects which are not close

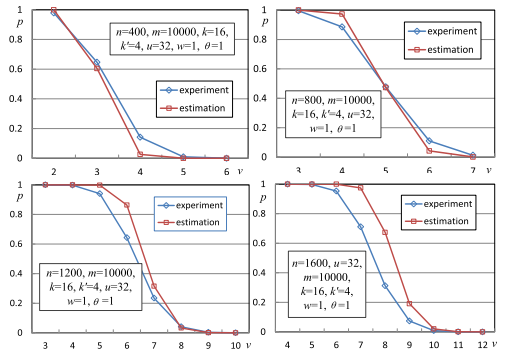

Fig. 5. FPRs with different number of objects.

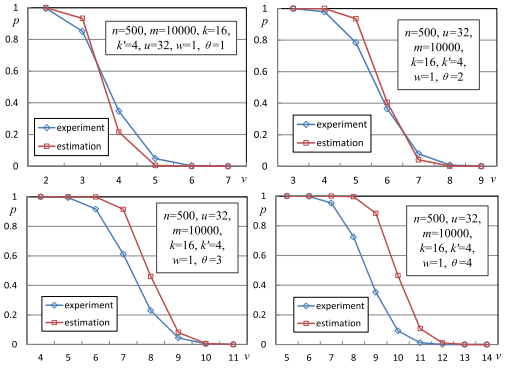

Fig. 6. FPRs with different granularities.

to any of the objects in the ILBF were used as test objects. Fig. 6 shows a similar pattern for the two curves from the theoretical estimation and experiments, respectively. As every θ SBFs are combined as a VSBF, the number of '1' increases with the increase of θ. Therefore, parameter υ should be set to large to control the false positive rate.

Theorem 3 may not give very accurate estimates because of the following assumptions: (1) we assume that the mapping locations (in the same SBF) for objects are not identical; (2) we assume that the k' events, i.e., k' counters are equal to or greater than the parameter υ, are independent. Unfortunately, similar to the discussion in [7], such events are slightly correlative. Nevertheless, Theorem 3 provides an FPR estimation for selected parameters.

An effective filter is desired to have both a low false positive rate and a low false negative rate. Comparing Fig.6 to Fig.3, we can choose a proper υ for each granularity θ to achieve the above goal. For example, we can choose $\upsilon = 6$, 9, 10, and 12 for $\theta = 1$, 2, 3, and 4, respectively. For each of the above cases, both the FPRs and the FNRs are very small.

We also compared the performance behaviors between the ILBF and the MLBF* using the above synthetic data. We notice that the computation cost mainly comes from multiply operations for the LSH functions adopted in the structures. To be fair, the number of LSH functions and the number of locations (i.e., the time and space costs) for both the ILBF and the MLBF* were set to be the same. The relevant parameters were as follows: for ILBF, we had $k = 16$, $u = 32$, $m = 8192$, $w = 1$; for MLBF*, we had $m' = 2^{18}$ (i.e., $32 * 8192$), $k' = 5$, $w = 1$, and either $k = 2$ and $L = 8$ or $k = 4$ and $L = 4$ (i.e., $kL = 16$). We stored 500 synthetic data

TABLE I

ILBF VERSUS MLBF* (SYNTHETIC DATA)

θ	MLBF* (k=2,L=8)		MLBF* (k=4,L=4)		ILBF(k=16)		
	FPR	FNR	FPR	FNR	FPR	FNR	υ
1	0.0014	0.0176	0.0000	0.4208	0.0032	0.0024	6
2	0.0066	0.0001	0.0000	0.0704	0.0008	0.0028	9
3*	0.0392	0.0000	0.0008	0.0064	0.0128	0.0000	10
4	0.0382	0.0000	0.0006	0.0092	0.0020	0.0054	12
5*	0.2478	0.0000	0.0069	0.0003	0.0026	0.0136	13
6*	0.2451	0.0000	0.0058	0.0005	0.0315	0.0051	13
7*	0.2432	0.0000	0.0046	0.0007	0.0114	0.0191	14
8	0.2388	0.0000	0.0041	0.0012	0.0846	0.0079	14
9*	0.8096	0.0000	0.0754	0.0003	0.0170	0.0609	15
10*	0.8015	0.0000	0.0736	0.0003	0.0679	0.0322	15

objects in each of the ILBF and the MLBF*. The difference in each dimension between the query object and one of the data object was randomly set to 0.100-0.145 and 10.0-14.5 to examine the FNRs and the FPRs, respectively.

As the MLBF* was designed for only filtering objects for multiple logarithmic granularities, the conservative larger logarithmic granularities were used to approximate the corresponding non-logarithmic granularities in TABLE I (i.e., marked with *). TABLE I shows that the MLBF* can control the both FPR and FNR to a low level for most of the granularities. The conservative method can dramatically increase FPRs, resulting in a "step-function" effect. Moreover, as mentioned in [6], adopting different numbers of hash functions in an MLBF* may only produce optimal results for some granularities (e.g., $k = 2$ is for $\theta = 1$, 2, while $k = 4$ is for $\theta = 4$, 8, from observing TABLE I), but not for others (e.g., poor FPR for $\theta = 4$, 8 when $k = 2$, and bad FNR for $\theta = 1$, 2 when $k = 4$, comparing to ILBF). On the other hand, for the ILBF structure, both the FPR and the FNR can be controlled by a proper parameter υ to a very low level for all the granularities (never exceed 9%). For the processing speed, both the ILBF and the MLBF* in the experiments spend about 0.011ms to filter an object.

V. CONCLUSION

Aiming for reducing transmission cost in network applications, we propose a novel filter structure ILBF, which can filter objects under multiple integer distance granularities. Theoretical analyses and experimental results show that the ILBF structure is efficient and effective. Extending the ILBF structure to allow an arbitrary (real number) distance granularity is an interesting open research issue.

REFERENCES

[1] A. Kirsch and M. Mitzenmacher, "Distance-sensitive Bloom filters," in *Proc. 8th Workshop Algorithm Eng. Experim.*, 2006, pp. 41–50.
[2] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.
[3] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.
[4] H. Lim, J. Lee, and C. Yim, "Complement Bloom filter for identifying true positiveness of a Bloom filter," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1905–1908, Nov. 2015.
[5] Y. Hua, B. Xiao, B. Veeravalli, and D. Feng, "Locality-sensitive Bloom filter for approximate membership query," *IEEE Trans. Comput.*, vol. 61, no. 6, pp. 817–830, Jun. 2012.
[6] J. Qian, Q. Zhu, and H. Chen, "Multi-granularity locality-sensitive Bloom filter," *IEEE Trans. Comput.*, vol. 64, no. 12, pp. 3500–3514, Dec. 2015.
[7] P. Bose *et al.*, "On the false-positive rate of Bloom filters," *Inf. Process. Lett.*, vol. 108, no. 4, pp. 210–213, Oct. 2008.