α Belief Propagation as Fully Factorized Approximation

First Hand Scientists
e-mail: doli@kth.se

Abstract

Belief propagation (BP) can do exact inference in loop-free graphs, but its performance could be poor in graphs with loops, and the understanding of its solution is limited. This work gives an interpretable belief propagation rule that is actually minimization of a localized α-divergence. We term this algorithm as α belief propagation (α-BP). The performance of α-BP is tested in MAP (maximum a posterior) inference problems, where α-BP can outperform (loopy) BP by a significant margin even in fully-connected graphs.

I. Introduction

Bayesian inference provides a general mathematical framework for many learning tasks such as classification, denoising, object detection, and signal detection. The wide applications include but not limited to imaging processing [1], multi-input-multi-output (MIMO) signal detection in digital communication [2], [3], inference on structured lattice [4], machine learning [5]–[7]. Specifically, statistic properties of a hidden variable $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ are of common interests in Bayesian inference. Practical interests usually include finding joint probability $p(\boldsymbol{x})$, marginal probability $p_i(x_i)$, the most probable state $\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x})$. It can be extended to maximum a posterior (MAP) inference when it is conditional on some observation ($\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\cdot)$). Direct inference from $p(\boldsymbol{x})$ may be difficult computationally or technically. For instance, in the MAP inference problem, it could be the case that the gradient or subgradient of $p(\boldsymbol{x})$ may not exist and it is computationally prohibitive to search $\boldsymbol{x}$'s whole feasible space.

Probabilistic graphical models as structured graphs provide a framework for modeling the dependency between random variables. Belief propagation (BP) is a general message-passing algorithm for performing inference on graphical models. The intuition of BP is exchange of belief (statistical information) between neighboring nodes [8]. When belief exchange converges, inference can be done by using the converged belief in graphical models. BP can solve inference problems exactly when the graphical model representation of $p(\boldsymbol{x})$ is loop-free or tree-structured [9]. When there are loops or circles in graphical models, BP is still a practical method to do inference approximately (loopy BP) by running it as if there is no loop. But its performance could be deteriorated significantly. In the loopy case, there are attempts to study convergence properties of BP in special cases [10], [11], but (loopy) BP may not converge in general.

Apart from the practical performance issues of BP in loopy graphs, the understanding of it is also limited. [12] shows that BP in loopy graphs approaches to a stationary point of an approximate free energy, the Bethe free energy in statistical physics. Based on this understanding, variants of BP are derived to improve BP. For instance, fractional BP in [13] applies a correction coefficient to each factor, generalized BP [12] propagates belief between different regions of a graph, and damping BP in [14] updates belief by combining old and new belief. Another track is expectation propagation (EP), introduced by Opper and Winther [15] and Minka [16], [17]. In EP, a simpler factorized distribution defined in exponential distribution family is used to approximate the original complex distribution, and an intuitive factor-wise refinement procedure is used to find such an approximate distribution. The method has an intuition of minimizing a localized Kullback-Leibler (KL) divergence. This is discussed further in [18] and it shows an unifying view of message passing algorithms. Following work stochastic EP [19] explores its variant method for applications to large dataset.

In this work, we take the path of Minka's variational methods to improve BP and also to gain better understanding of BP in loopy graphs. We define a surrogate distribution $q(\boldsymbol{x})$ first. $q(\boldsymbol{x})$ is assumed to be fully factorized and each factor of $q(\boldsymbol{x})$ represents a message in the factor graph representation of the original distribution $p(\boldsymbol{x})$. Fully factorization is the only requirement to $q(\boldsymbol{x})$. Then we define a message passing rule that is derived by minimizing a localized α-divergence. This is factor-wise refinement of $q(\boldsymbol{x})$ iteratively. We refer to the obtained algorithm by α-BP. The merits of α are as follows:

a. α-BP has clear intuition as localized minimization of α-divergence between original distribution p and surrogate distribution q.
b. α-BP generalizes the standard BP, since the message rule of BP is a special case of α-BP.
c. α-BP could outperform BP significantly even in full-connected graphs while still maintaining simplicity of BP for inference.

II. Preliminary

In this section, we provide the preliminaries that are needed in this paper. We introduce the α-divergence and a graphical model that we are going to use to explain α-BP.

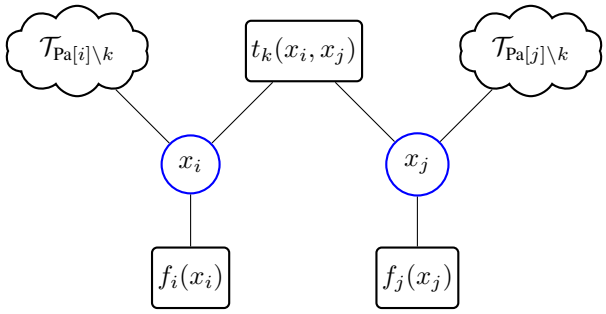

Figure 1. Factor graph illustration of Equation 5.

A. Divergence Measures

As explained in Section I, we are going to minimize α-divergence between p and q, which is defined as follows according to [20] [18]:

$$\mathcal{D}_\alpha(p\|q) = \frac{\int_{\boldsymbol{x}} \alpha p(\boldsymbol{x}) + (1-\alpha)q(\boldsymbol{x}) - p(\boldsymbol{x})^\alpha q(\boldsymbol{x})^{1-\alpha} d\boldsymbol{x}}{\alpha(1-\alpha)}, \tag{1}$$

where α is the parameter of α-divergence, distribution p and q are unnormalized, i.e. $\int_{\boldsymbol{x}} p(\boldsymbol{x})d\boldsymbol{x} \neq 1$, $\int_{\boldsymbol{x}} q(\boldsymbol{x})d\boldsymbol{x} \neq 1$.

The classic KL divergence is defined as

$$KL(p\|q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} + \int q(\boldsymbol{x}) - p(\boldsymbol{x}) d\boldsymbol{x} \tag{2}$$

where the $\int q(\boldsymbol{x}) - p(\boldsymbol{x})d\boldsymbol{x}$ is a correction factor to accommodate unnormalized p and q. The KL divergence is a special case of α-divergence, since $\lim_{\alpha\to 1} \mathcal{D}_\alpha(p\|q) = KL(p\|q)$ and $\lim_{\alpha\to 0} \mathcal{D}_\alpha(p\|q) = KL(q\|p)$, by applying L'Hôpital's rule to Equation 1.

Both α-divergence and KL divergence are equal to zero if $p = q$, and they are non-negative (therefore satisfy the basic property of error measure). Denote KL-projection by

$$\text{proj}[p] = \operatorname*{argmin}_{q\in\mathcal{F}} KL(p\|q), \tag{3}$$

where $\mathcal{F}$ is the distribution family of q.

According to the stationary point equivalence Theorem in [18], $\text{proj}[p^\alpha q^{1-\alpha}]$ and $\mathcal{D}_\alpha(p\|q)$ have same stationary points. A heuristic scheme to find q minimizing $\mathcal{D}_\alpha(p\|q)$ is to find its stationary point by a fixed-point iteration:

$$q(\boldsymbol{x})^{\text{new}} = \text{proj}[p(\boldsymbol{x})^\alpha q(\boldsymbol{x})^{1-\alpha}]. \tag{4}$$

B. A Graphic Model

We introduce a pairwise Markov random field (MRF) $p(\boldsymbol{x})$ to explain our algorithm. Variable $\boldsymbol{x} := \{x_1, x_2, \cdots, x_N\}$, where $x_i \in \mathcal{A}$ is a discrete finite set or subset of $\mathbb{R}$, and N is a positive integer. An undirected graphical model, known as Markov random field, defines a family of joint probability distributions over $\boldsymbol{x}$ by associating its index set with the vertex set $\mathcal{V}$ of undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$. The graph contains edges $\mathcal{E} \subset \mathcal{V}\times\mathcal{V}$, where a pair of $(v, u) \in \mathcal{E}$ if and only if nodes v and u are connected by a edge. For a clique $\mathcal{C}$ of the graph, let $\varphi_{\mathcal{C}} : \mathcal{A}^{|\mathcal{C}|} \to (0, \infty)$ be a potential function of vector $\boldsymbol{x}_{\mathcal{C}} := \{x_v | v \in \mathcal{C}\}$. Let us then write the joint distribution of $\boldsymbol{x}$ as

$$p(\boldsymbol{x}) \propto \prod_{v\in\mathcal{V}} \varphi_v(x_) \prod_{(v,u)\in\mathcal{E}} \varphi_{vu}(x_v, x_u), \tag{5}$$

where $\propto$ denotes the fact that the only difference between two sides of $\propto$ is a constant factor.

The factor graph representing Equation 5 is shown in Figure 1. In the figure, $\text{Pa}[i]$ is the index set of pairwise factors connecting to variable node x_i, i.e. $\text{Pa}[i]$ is subset of $\mathcal{K}$, $\setminus$ denotes exclusion. $\mathcal{T}_{\text{Pa}[i]\setminus k}$ is the product of all pairwise factors connecting to x_i except for t_k:

$$\mathcal{T}_{\text{Pa}[i]\setminus k} = \prod_{n\in\text{Pa}[i]\setminus k} t_n. \tag{6}$$

III. α-BP as Fully-Factorized Approximation

In this section, we will show why α-BP as a message-passing algorithm can be used as a fully-factorized approximation to the original distribution $p(\boldsymbol{x})$.

A. α Belief Propagation

1) Fully Factorized Surrogate: Now we formulate a surrogate distribution as

$$q(\boldsymbol{x}) \propto \prod_{v\in\mathcal{V}} \tilde{\varphi}_v(x_) \prod_{(v,u)\in\mathcal{E}} \tilde{\varphi}_{vu}(x_v, x_u), \tag{7}$$

to approximate $p(\boldsymbol{x})$. The surrogate distribution would be used to estimate inference problems of $p(\boldsymbol{x})$. We further assume that $q(\boldsymbol{x})$ can be fully factorized, which means that $\tilde{\varphi}_{v,u}(x_v, x_u)$ can be factorized as two independent functions of x_v, x_u respectively. We denote this factorization as

$$\tilde{\varphi}_{v,u}(x_v, x_u) := m_{uv}(x_v) m_{vu}(x_u). \tag{8}$$

We use the notation $m_{uv}(x_v)$ to denote the factor as a function of x_v due to the intuitive fact that m_{uv} is also the message from the factor $\varphi_{uv}(x_u, x_v)$ to variable node x_v. Similarly we have factor $m_{vu}(x_u)$. Then the marginal can be formulated straightforwardly as

$$q_v(x_v) \propto \tilde{\varphi}_v(x_v) \prod_{w\in\mathcal{N}(v)} m_{wv}(x_v). \tag{9}$$

2) Local α-Divergence Minimization: Now, we are going to use the heuristic scheme as in Equation 4 to minimize the information loss by using tractable $q(\boldsymbol{x})$ to represent $p(\boldsymbol{x})$. The information loss is measured by α-divergence $\mathcal{D}_\alpha(p(\boldsymbol{x})\|q(\boldsymbol{x}))$.

We do factor-wise refinement to update the factors of $q(\boldsymbol{x})$ such that $q(\boldsymbol{x})$ approaches $p(\boldsymbol{x})$ asymptotically similar to [16], [18]. Without losing generality, we begin to refine factor $\tilde{\varphi}_{uv}(x_u, x_v)$. Define $q^{\backslash(u,v)}(\boldsymbol{x})$ as all other factors except for $\tilde{\varphi}_{uv}(x_u, x_v)$

$$q^{\backslash k}(\boldsymbol{x}) = q(\boldsymbol{x})/\tilde{\varphi}_{uv}(x_u, x_v) \propto \prod_{v\in\mathcal{V}} \tilde{\varphi}_v(x_) \prod_{(v,u)\in\mathcal{E}\backslash(u,v)} \tilde{\varphi}_{vu}(x_v, x_u). \tag{10}$$

Similarly, we have $p^{\backslash(u,v)}(\boldsymbol{x})$ as all other factors except for $\varphi_{uv}(x_u, x_v)$. Assume that we already have had $q^{\backslash(u,v)}(\boldsymbol{x})$ as a good approximation of $p^{\backslash(u,v)}(\boldsymbol{x})$, i.e. $q^{\backslash(u,v)}(\boldsymbol{x}) \simeq p^{\backslash(u,v)}(\boldsymbol{x})$, it is $\tilde{\varphi}_{uv}(x_u, x_v)$ that remains to be refined. Then the problem $\operatorname{argmin}_{\tilde{\varphi}^{\text{new}}_{uv}} \mathcal{D}_\alpha\left(p^{\backslash(uv)}\varphi_{uv}\|q^{\backslash(u,v)}\tilde{\varphi}^{\text{new}}_{uv}\right)$ becomes

$$\operatorname*{argmin}_{\tilde{\varphi}^{\text{new}}_{uv}(x_u,x_v)} \mathcal{D}_\alpha\left(q^{\backslash(u,v)}(\boldsymbol{x})\varphi_{uv}(x_u, x_v)\|q^{\backslash(u,v)}(\boldsymbol{x})\tilde{\varphi}^{\text{new}}_{uv}(x_u, x_v)\right), \tag{11}$$

which searches for new factor $\tilde{\varphi}^{\text{new}}_{uv}(x_u, x_v)$ such the above divergence is minimized. Using Equation 4, the above problem is equivalent to

$$\begin{aligned} & q^{\backslash(u,v)}(\boldsymbol{x})\tilde{\varphi}^{\text{new}}_{uv}(x_u, x_v) \\ & \propto \text{proj}\left[\left(q^{\backslash(u,v)}(\boldsymbol{x})\varphi_{uv}(x_u, x_v)\right)^\alpha \left(q^{\backslash(u,v)}(\boldsymbol{x})\tilde{\varphi}_{uv}(x_u, x_v)\right)^{1-\alpha}\right] \\ & \propto \text{proj}\left[q^{\backslash(u,v)}(\boldsymbol{x})\varphi_{uv}(x_u, x_v)^\alpha \tilde{\varphi}_{uv}(x_u, x_v)^{1-\alpha}\right]. \end{aligned} \tag{12}$$

Let us refine one message per time on edge (u, v). Without lose of generality, we update m_{uv} and denote

$$\tilde{\varphi}^{\text{new}}_{uv}(x_u, x_v) = m^{\text{new}}_{uv}(x_v) m_{vu}(x_u). \tag{13}$$

Since KL-projection to a fully factorized distribution reduces to matching the marginals, Equation 12 is reduced to

$$\sum_{\boldsymbol{x}\backslash x_v} q^{\backslash(u,v)}(\boldsymbol{x})\tilde{\varphi}^{\text{new}}_{uv}(x_u, x_v) \propto \sum_{\boldsymbol{x}\backslash x_v} q^{\backslash(u,v)}(\boldsymbol{x})\varphi_{uv}(x_u, x_v)^\alpha \tilde{\varphi}_{uv}(x_u, x_v)^{1-\alpha}. \tag{14}$$

We use summation here. But it should be replaced by integral if $\mathcal{A}$ is a continuous set. Solving Equation 14 gives the message passing rule as

$$m^{\text{new}}_{uv}(x_v) \propto m_{uv}(x_v)^{1-\alpha}\left[\sum_{x_u} \varphi_{uv}(x_u, x_v)^\alpha m_{vu}(x_u)^{1-\alpha} \tilde{\varphi}_u(x_u) \prod_{w\in\mathcal{N}(u)\backslash v} \varphi_{wu}(x_w, x_u)\right]. \tag{15}$$

As for the singleton factor $\tilde{\varphi}_v(x_v)$, we can do the refinement procedure on $\tilde{\varphi}_v(x_v)$ in the same way as we have done for $\tilde{\varphi}_{uv}(x_u, x_v)$. This gives us the update rule of $\tilde{\varphi}_v(x_v)$ as

$$\tilde{\varphi}^{\text{new}}_v(x_v) \propto \varphi_v(x_v)^\alpha \tilde{\varphi}_v(x_v)^{1-\alpha}, \tag{16}$$

which is the belief from factor $f_v(x_v)$ to variable x_v. Note, if we initialize $\tilde{\varphi}_v(x_v) = \varphi_v(x_v)$, then it remains the same in all iterations, which makes

$$m^{\text{new}}_{uv}(x_v) \propto m_{uv}(x_v)^{1-\alpha}\left[\sum_{x_u} \varphi_{uv}(x_u, x_v)^\alpha m_{vu}(x_u)^{1-\alpha} \varphi_u(x_u) \prod_{w\in\mathcal{N}(u)\backslash v} \varphi_{wu}(x_w, x_u)\right]. \tag{17}$$

B. Remarks on α-BP

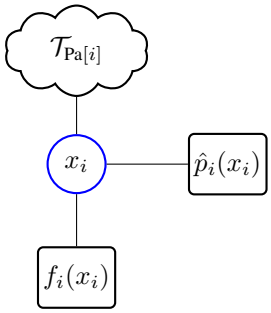

Figure 2. Factor graph illustration with prior factor.

As discussed in Section II, $KL(p\|q)$ is the special case of $\mathcal{D}_\alpha(p\|q)$ when $\alpha \to 1$. When applying $\alpha = 1$ to Equation 17, it gives

$$m_{uv}^{\text{new}}(x_v) \propto \left[\sum_{x_u} \varphi_{uv}(x_u, x_v)\varphi_u(x_u) \prod_{w \in \mathcal{N}(u)\setminus v} \varphi_{wu}(x_w, x_u) \right], \tag{18}$$

which is exactly the messages of BP in Chapter 8 of [8]. From this point of view, α-BP generalizes BP.

Inspired by [21] and assembling methods [22], we can add an extra singleton factor to each x_i as prior information that is obtained from other (usually weak) methods. This factor stands for our belief from exterior estimation. Then run our α-BP. Denote the prior by $\hat{p}_i(x_i)$ for variable node x_i, then the factor graph including this prior belief can be represented as in Figure 2.

We summarize the method into the pseudo-code in Algorithm 1. Though we explain the method with a binary MRF, it is straightforward to replace the factor t_k by a factor involving more than two variables and applies α-BP to general factor graphs.

Algorithm 1 Algorithm of α-BP

Input: Factor graph of $p(\boldsymbol{x})$
1: Initialize $q(\boldsymbol{x})$
2: **if** Prior belief on x_i available **then**
3: Add prior factor as Figure 2
4: **end if**
5: **while** not converge **do**
6: **for** each edge of factor graph **do**
7: Message update by Equation 17 or Equation 16
8: **end for**
9: **end while**
10: **return** $q(\boldsymbol{x})$

References

[1] R. Zhang, C. A. Bouman, J. Thibault, and K. D. Sauer, "Gaussian mixture markov random field for image denoising and reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*, Dec 2013, pp. 1089–1092.

[2] J. Cspedes, P. M. Olmos, M. Snchez-Fernndez, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional mimo systems," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2840–2849, Aug 2014.

[3] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large mimo detection via approximate message passing," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1227–1231.

[4] N. Friel, A. N. Pettitt, R. Reeves, and E. Wit, "Bayesian inference in hidden markov random fields for binary data defined on large lattices," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 243–261, 2009. [Online]. Available: http://www.jstor.org/stable/25651244

[5] G. Montufar, "Restricted Boltzmann Machines: Introduction and Review," *ArXiv e-prints*, Jun. 2018.

[6] G. Lin, C. Shen, I. Reid, and A. v. d. Hengel, "Deeply learning the messages in message passing inference," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 361–369. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969280

[7] K. Yoon, R. Liao, Y. Xiong, L. Zhang, E. Fetaya, R. Urtasun, R. S. Zemel, and X. Pitkow, "Inference in probabilistic graphical models by graph neural networks," *CoRR*, vol. abs/1803.07710, 2018. [Online]. Available: http://arxiv.org/abs/1803.07710

[8] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[9] F. R. Kschischang, B. J. Frey, and H. . Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, Feb 2001.

[10] A. T. Ihler, J. W. Fischer III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *J. Mach. Learn. Res.*, vol. 6, pp. 905–936, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088703

[11] J. Du, S. Ma, Y.-C. Wu, S. Kar, and J. M. F. Moura, "Convergence analysis of distributed inference with vector-valued gaussian belief propagation," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6302–6339, Jan. 2017. [Online]. Available: http://dl.acm.org/citation.cfm?id=3122009.3242029

[12] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, ser. NIPS'00. Cambridge, MA, USA: MIT Press, 2000, pp. 668–674. [Online]. Available: http://dl.acm.org/citation.cfm?id=3008751.3008848

[13] W. Wiegerinck and T. Heskes, "Fractional belief propagation," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, ser. NIPS'02. Cambridge, MA, USA: MIT Press, 2002, pp. 438–445. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968618.2968673

[14] M. Pretti, "A message-passing algorithm with damping," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11, pp. P11 008–P11 008, nov 2005. [Online]. Available: https://doi.org/10.1088%2F1742-5468%2F2005%2F11%2Fp11008

[15] M. Opper and O. Winther, "Gaussian processes for classification: Mean-field algorithms," *Neural Comput.*, vol. 12, no. 11, pp. 2655–2684, Nov. 2000. [Online]. Available: http://dx.doi.org/10.1162/089976600300014881

[16] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369. [Online]. Available: http://dl.acm.org/citation.cfm?id=647235.720257

[17] ——, "A family of algorithms for approximate bayesian inference," Ph.D. dissertation, Cambridge, MA, USA, 2001, aAI0803033.

[18] T. Minka, "Divergence measures and message passing," Tech. Rep. MSR-TR-2005-173, January 2005. [Online]. Available: https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/

[19] Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "Stochastic expectation propagation," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2323–2331. [Online]. Available: http://papers.nips.cc/paper/5760-stochastic-expectation-propagation.pdf

[20] H. Zhu and R. Rohwer, "Information geometric measurements of generalisation," Tech. Rep., 1995.

[21] J. Goldberger and A. Leshem, "Pseudo prior belief propagation for densely connected discrete graphs," in *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, Jan 2010, pp. 1–5.

[22] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[23] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.