Theory of metal-semiconductor transitions in random impurity bands

J. C. Phillips
AT&T Bell Laboratories, Murray Hill, New Jersey 07974
(Received 15 April 1991)

In Si:P a metal-semiconductor transition occurs homogeneously with randomly distributed impurities. Neither classical percolation theory nor classical scaling theory describes this transition, but it is explained by set theory and the quantum theory of measurement. These concepts explain all aspects of the conductivity transition as well as the absence of accompanying transitions in the specific heat and the magnetic susceptibility.

I. INTRODUCTION

The quantum structure of electronic systems undergoing metal-semiconductor transitions has been a challenging theoretical problem for more than four decades. The pioneering papers of Mott[1-3] and Anderson[4] stimulated many experiments and led to the recognition that ultrapure semiconductors (such as Si) doped with shallow impurities (such as P) represent the physical system best suited for studying quantum effects. These are fully resolved, however, only at ultralow temperatures $10^{-3} \lesssim T < 10^{-2}$ K, even though the equivalent Fermi temperature at the critical concentration n_c in Si:P is $T_c \sim 100$ K. These conditions were finally realized in classic experiments carried out by Thomas, Paalanen, and Rosenbaum[5] (TPR) at ^{3}He temperatures using uniaxial strain to vary $n - n_c$ through the transition in a single sample.

Prior to the TPR experiment there was a wide range of theoretical opinions concerning the behavior of $\sigma_T(n - n_c)$ in the limit $T \to 0$, as described by the exponent α in the relation

$$\sigma_0(n - n_c) = b\sigma_{\rm IR}[(n - n_c)/n_c]^{\alpha}, \quad (1)$$

where $\sigma_{\rm IR}$, the Ioffe-Regel conductivity, is calculated[5] for an electron gas with an appropriate effective mass and a mean free path l equal to the average impurity spacing d. (Here b is a numerical constant of order unity which includes the effects of orbital valley degeneracy.) For the most part Mott[1-3] had argued that σ would be discontinuous at $n = n_c$, corresponding to $\alpha = 0$ and a first-order phase transition. Very early, in the different context of amorphous semiconductors, Cohen had suggested[6] that $\alpha \sim \frac{1}{2}$. A classical scaling argument by Abrahams *et al.*[7] had led to $\alpha \sim 1$. Many other approaches, including a hydrodynamic mode-coupling[8] model ($\alpha = \frac{1}{2}$) and a "rigorous" field-theoretic[9] model ($\alpha \geq 2/3$), were also available. None of these models predicted the temperature dependence of $\sigma_T(n - n_c)$ for n near n_c, or the crossover compensation level[10] K_c at which α might change rapidly as the level K of compensating acceptors is altered.

One might have expected that the high level of theoretical effort which preceded the TPR experiment would have increased further after their definitive results, but the opposite actually seems to have been the case. Broadly speaking, three approaches have subsequently been adopted to explain the available data: (1) the interaction description, as developed by Lee and co-workers;[11,12] the set-theoretic method,[13-15] which is the subject of the present paper; and numerical simulations.[16,17] The latter are generally carried out on simplified models which appear to contain the same essential features as Si:P. However, in disordered systems meaningful numerical results can be achieved only at high temperatures or short times, and one must then demonstrate that as T is lowered (in this case to $T \lesssim 10^{-4} T_F$) the statistical samples remain in equilibrium. If this procedure is not used, then certain simplifying assumptions must be made, and the validity of these assumptions can be tested only by comparison with the TPR experiment. So formidable are the statistical problems at ultralow temperatures or long times that such comparisons are usually inconclusive.

Recently I have become convinced that the set-theoretic method[13-15] contains all the features needed to explain the TPR experimental data. The purpose of this paper is to provide a full discussion of these features, previously described only briefly. This discussion is necessary because it involves deep problems in set theory (the axiom of choice as discussed by Gödel and Cohen[18]) as well as the quantum theory of measurement (as discussed by Einstein[19] and Bohr[20]). The set-theoretic method not only explains the behavior of $\sigma(n, T, K)$, but it also enables us to understand why the abrupt phase transition in σ is accompanied only by broad transitions[21,22] in the specific heat c and the magnetic susceptibility χ.

II. SET-THEORETIC METHOD

Abrupt behavior (σ_0 or $d\sigma_0/dn$ discontinuous) of $\sigma_0(n - n_c)$ implies a phase transition from localized to extended states at $E \simeq E_F$. The phase transition in turn is associated with a large system of N electrons in the limit $N \to \infty$, so that we are dealing with an infinite set of electronic states which comprise a continuum. The procedures for handling sets of infinite states, and for separating these infinite sets into infinite subsets, are described by axiomatic set theory, and in particular by the axiom of choice, which assumes that infinite sets can be

 ©1992 The American Physical Society

separated uniquely into infinite subsets even when no explicit algorithm for doing so is known. [The best-known mathematical example is the theorem that real numbers, like integers, form a well-ordered (alphabetizable) set.[23]] This axiom is independent of the other axioms of set theory, and in particular cases its applicability is an open question. This means that the separability of localized and extended states must be decided by experiment, and cannot be determined *a priori,* for instance, by saying that it is not possible because no explicit procedure for doing so is known.

The set-theoretic method[24] is different from other methods of solving analytically intractable problems, for instance, perturbation theory, group theory, or scaling theory. Like scaling theory it does not rely on explicit solutions, but unlike scaling theory (which deals with correlation functions, i.e., classical properties) it can treat quantum-mechanical properties associated with both phases and amplitudes of wave functions. Here it relies on analogies with simpler examples of finite (usually small) systems of wave functions, or on analogies with wave functions in ordered systems. Obviously these analogies must be handled with care, and the ultimate judge is experiment.

Sometimes a set-theoretic method has been adopted intuitively, without explicit recognition of its distinctive character. The classic example in the theory of conduction of random impurity bands is Mott's celebrated solution for $n < n_c$, that is, variable-range hopping in the insulating regime.[25] Previously, it was assumed that the conduction was percolative, with nearest-neighbor hopping, but Mott showed that the hopping range in d dimensions increases like $T^{-[1/(d+1)]}$ as $T \rightarrow 0$. His derivation relies on optimizing this range with respect to the *set* of all possible localized states within an energy range of order $W \sim [R^d N(E_F)]^{-1}$ of E_F with $W \sim kT$. His solution is inherently nonperturbative and nonclassical.

Because all states are localized for $n < n_c$ there the issue of the existence of extended states and their separability from localized states is not a problem. For $n > n_c$ Mott cut the Gordian knot and chose an equally simple solution, namely localized states for $E < E_0$ and extended states for $E > E_0$. This gives $\alpha = 0$ in (1), and for a long time it appeared that such discontinuous behavior was compatible with experiment. However, with improved cryogenics, enabling measurements closer to $T = 0$, and higher quality samples, it gradually became clear that the transition was probably continuous, $\alpha > 0$. The actual value $\alpha = \frac{1}{2}$ was finally obtained in the TPR experiments, and this revived the basic question of coexistence of localized and extended states at the same energy, and their separability.

III. QUANTUM THEORY OF MEASUREMENT IN HOMOGENEOUS DISORDERED SYSTEMS

Results of quantum experiments may be paradoxical when interpreted in classical terms,[19] but the paradoxes are resolved when the effects of the measurement process itself are taken into account.[20] The measurement process acts as a kind of projection operator which separates all possible correlations of the system variables into two sets, those statistically consistent with the measurement and those not. In the case of only $N = 2$ particles fifteen parameters must be measured to determine all possible correlations, and this number rapidly grows unmanageable as N increases.[19] For a quantum phase transition it appears that set-theoretic methods must be employed when the disorder in the system is so great that thermodynamic parameters of the Landau type can no longer be defined.

The first step in applying set-theoretic methods to the metal-semiconductor transition in random impurity bands is the recognition that the critical dimensionality d_c for the existence of extended states can be determined by using the uncertainty principle[13] to map this problem onto the classical model of a random-field Ising system.[26] The presence of electrodes spaced a distance L apart means that there is an energy uncertainty

$$\delta E = (d/L) E_I \ , \tag{2}$$

where d is the average impurity spacing and E_I is the impurity ionization energy. The number of states N_δ in this energy interval scales as

$$N_\delta \sim L^{d-1} \ , \tag{3}$$

while the disordering effects produce localized states whose number N_l scales as

$$N_l \sim L^{d/2} \ . \tag{4}$$

Comparing (3) and (4) we see that the critical dimensionality above which extended states can exist is d_c, where

$$d_c - 1 = d_c/2 \tag{5}$$

or

$$d_c = 2 \ . \tag{6}$$

This derivation is mathematically isomorphous to that used to derive d_c in the random-field Ising model in a seminal paper.[26] Before the derivation of d_c according to Eqs. (3)–(6), the same value had been derived heuristically (without using the uncertainty principle) by classical scaling.[7]

This derivation is not only important in its own right, but it also contains the kernel of the procedure which is required to calculate α. The energy interval (2) refers to a set of states measured with a certain orientation of the electric field $\mathbf{F}$. The extended states which exist for $d > d_c$ and $E > E_0$ must reduce to amplitude-modulated but phase-coherent states similar to plane-wave states described by wave numbers $k_\parallel$, where

$$k_\parallel = \mathbf{k} \cdot \mathbf{F}/F \ . \tag{7}$$

In other words, the applied field separates the extended states from the localized states only parallel to the applied field. The total number of extended states in d dimensions N_e from $\mathbf{k} = 0$ to $k = k_\parallel = k_\perp = k_F$ is propor-

tional to

$$N_e(\mathbf{k}) \sim k_{\parallel} k_{\perp}^{d-1} , \tag{8}$$

but only $k_{\parallel}$ is defined by the measurement process. Because

$$E - E_0 \sim k^2 , \tag{9}$$

this means that the conductivity $\sigma(E_F - E_0)$ is given by

$$\sigma_0 \sim N\tau . \tag{10}$$

For E_F near E_0 the number of extended states $N_e(\mathbf{k})$ is much smaller than the number of localized states $N_l(E_F)$. Then τ^{-1}, which describes the scattering rate of extended states into both extended and localized states, is approximately constant and

$$\sigma_0 \sim (N_e)^{1/d}(N_e + N_l)^{(d-1)/d} \tag{11}$$

because only parallel to the field have the extended states been separated from the localized states by the applied field. In other words, the factor $k_{\perp}$ in (8) is replaced by $(N_e + N_l)^{1/d}$, which is nearly constant. This leaves

$$\sigma_0 \sim N_e^{1/d} \sim (E_F - E_0)^{1/2} \tag{12}$$

and with

$$E_F - E_0 \sim n - n_c \tag{13}$$

we obtain

$$\alpha = \tfrac{1}{2} \tag{14}$$

in (1), in agreement with experiment.

What is the physical meaning of a percolative path which is described by extended (phase-coherent) states with a density $N_e(k_{\parallel}) \sim k_{\parallel}$ parallel to the field and a nearly constant density perpendicular to the field? Since this nearly constant density consists almost entirely of localized states, we can imagine that these states are localized on a scale of order a, where a is the average impurity spacing, and that the percolation threshold, $N_e(E_F) \to 0$ as $E_F \to E_0$, arises entirely because $k_{\parallel} \to 0$ as $E_F \to E_0$, i.e., only increasingly long-wavelength states which oscillate parallel to the applied field **F** are extended as $E_F \to E_0$. This model can be described as quantum percolation; it is qualitatively completely different from classical percolation, which sees fewer classical paths crossing planes transverse to **F** as $E_F \to E_0$ and which predicts[27] $\alpha \sim 1.6$ for $d = 3$. It seems to me that this model represents the natural generalization of Mott's set-theoretic derivation of variable-range hopping for the semiconductive regime $n < n_c$ to the metallic regime $n > n_c$, but augmented by the state selectivity implicit in the quantum theory of measurement.

Let us revert now to the set-theoretic issues discussed in Sec. II. One could interpret the factor $(N_e + N_l)^{(d-1)/d}$ in Eq. (11) as evidence for inertial drag of localized and extended states transverse to **F** by the phase-coherent motion of the extended states parallel to **F**. Alternatively, the set-theoretic interpretation is that the projective effect of the measurement process has enabled us to implement separability via the axion of choice parallel to **F**, but that it has had no effect transverse to **F**. This lack of effectiveness transverse to **F** was not discussed explicitly by me in earlier papers[13] and it is the central point emphasized here. In other words, in the present context we have *partial* separability with respect to establishing the longitudinal but not transverse phases of quasiparticle states.

IV. THERMAL ACTIVATION

With increasing temperature T the localized states $N_l(E_F)$ increasingly contribute[5] to $\sigma_T(n - n_c)$, as shown for the reader's convenience in Fig. 1. These states are activated by thermal number fluctuations which enhance $N(E_F)$ above the value

$$N_\sigma = \{N_e(E_F)[N_e(E_F) + N_l(E_F)]^{d-1}\}^{1/d} \tag{15}$$

appropriate to $T = 0$. Physically we imagine that in addition to the phase-coherent states at $T = 0$ there are many other states which would be extended but whose phase coherence is broken at a few weak links.[14] These states are made conductive by thermal fluctuations which locally increase $N(E_F)$. The general relation[28] for such number fluctuations is

$$\overline{(\Delta N)^2} = kT \left[\frac{\partial N}{\partial \mu} \right]_{T,V} , \tag{16}$$

where μ is the chemical potential; for hydrogenic impurities $(\partial N/\partial \mu)_{T,V} = E_I^{-1}$, where E_I is the impurity binding energy (540 K for Si:P, compared to $T_F \sim 100$ K). Be-

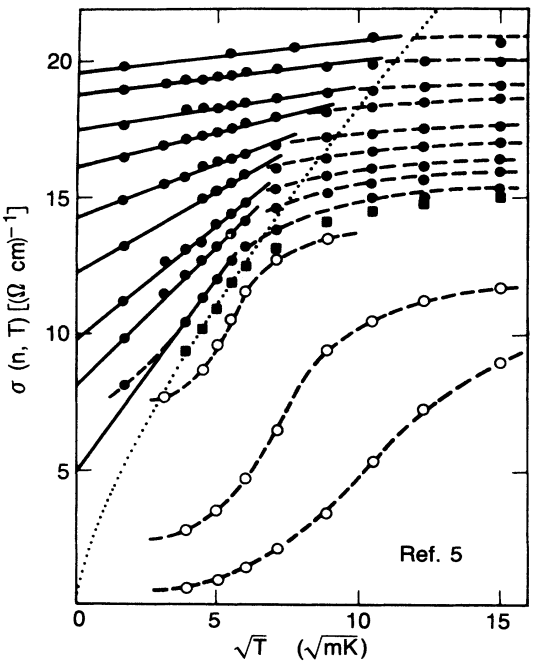

FIG. 1. Electrical conductivity $\sigma(n,T)$ for Si:P for a range of values of $0 < (n - n_c)/n_c \lesssim 10^{-2}$, as measured in Ref. 5. The solid lines are fits over regions where $\Delta\sigma \propto T^{1/2}$.

cause $\sigma \propto N$, we need

$$\langle \Delta N \rangle \equiv [\overline{(\Delta N)^2}]^{1/2} = (kT/E_I)^{1/2} , \quad (17)$$

which means that

$$\delta\sigma = \sigma_T - \sigma_0 = g(n)\sigma_{IR}(kT/E_I)^{1/2} , \quad (18)$$

where $g(n)$ is a geometrical factor describing the relative lengths of insulating and conductive regions. The relevant length scales for fluctuations are a and l_{ex} [the mean free path for exchange scattering with states localized on background (electrically inactive) impurities, such as deep donors]. Comparison with the experimental data, shown in Fig. 1, confirms the $T^{1/2}$ behavior of (17) up to $T \sim 0.1$ K and shows that $g(n_c) \sim 10$, while $g(1.01n_c) \sim 1$, which suggests that $l_{ex} \sim 10a$.

We may note that we know little about the geometry of the thermally excited states, but at the same time Eq. (17) contains no length scales. This relation is valid so long as $kT \ll E_I$, which is the case here. With increasing temperature the fluctuations in N may overlap in such a way that their contributions to $\delta\sigma$ are redundant. This would explain the leveling off of $\delta\sigma$ for $T \gtrsim 0.1$ K.

V. COMPENSATION AND CROSSOVER TO CLASSICAL PERCOLATION

Because the exponent $\alpha = 0.5$ in Si:P depends on the phase coherence of extended states formed as linear combinations of states in the energy range $E_F \pm \delta E$, where δE is given by Eq. (2), it seems very likely that exchange between electrons in localized states associated with other impurities (acceptors or deep donors, for example) can destroy this coherence, causing the transition to revert to a classical form with $\alpha \gtrsim 1$. Unfortunately few data are available for Si:P, and the closest approximation (in terms of shallow donors) is Ge:Sb. These data[10] lie in a different range than those shown in Fig. 1. Using Mott's σ_{min} for the conductivity scale,[10] with $\sigma_{min} = 20\ (\Omega\,\mathrm{cm})^{-1}$ in Si:P and $7\ (\Omega\,\mathrm{cm})^{-1}$ in Ge:Sb, all the data in Fig. 1 lie below σ_{min}, while the data for Ge:Sb lie between σ_{min} and $15\sigma_{min}$. Moreover, the data shown in Fig. 1 correspond to $n/n_c - 1 \lesssim 10^{-2}$, while the Ge:Sb data span the range $0.2 < n/n_c - 1 < 4$. Thus the Ge:Sb data are not in the critical range where quantum effects are dominant. However, the Ge:Sb data do give an indication of a crossover in α for a compensation level $K_c \lesssim 0.05$ with $\alpha \sim 0.7$, and we wish to interpret this indication in the context of the crossover from classical to quantum percolation.

Classical percolation is pictured in the context of nearest-neighbor hopping and on lattices with coordination numbers Z and with sites occupied with a probability p, the percolation threshold p_c (or critical density n_c) typically corresponds to an average number $\bar{N}$ of occupied nearest sites with $\bar{N}_c = p_c Z$ between 2 and 3.[27] For Si:P with n near n_c we can imagine in three dimensions that $Z \sim 10$ (as one would have for random sphere packing). A coherent wave packet for $n - n_c \gtrsim 0.1n_c$ should then involve a superposition of $P(1s, 2s, 2p)$ states along a path uninterrupted by exchange. Such a path should involve only P donors which contain no acceptor atoms in their near neighbor sphere; these donors can be called "pure" donors, as distinguished from compensated donors which contain one or more acceptor atoms as near neighbors. The fraction f_0 of pure donors is given by Poisson statistics[28] as

$$f_0 = e^{-N_1 K} , \quad (19)$$

where the compensation level is

$$K = n_A/n_D . \quad (20)$$

Next we calculate the number of pure donors N_0 in the near-neighbor shell of a pure donor,

$$N_0 = Zf_0 , \quad (21)$$

and near the crossover compensation level K_c

$$N_0 = 2.5(5) , \quad (22)$$

which with $Z = 10$ gives

$$K_c = 0.14(2) . \quad (23)$$

By interpolating linearly on (K, α) between (0,0.5) and (0.14,1), we obtain a crossover near

$$(K_1, \alpha_1) = (0.06(1), 0.7) , \quad (24)$$

in excellent agreement with the sparse experimental data[10] with $K_1 \lesssim 0.05$.

VI. SPECIFIC HEAT AND MAGNETIC SUSCEPTIBILITY

The specific heat and its magnetic-field dependence, as well as the magnetic susceptibility, have been studied[22] down to $T \gtrsim 30$ mK in both compensated and uncompensated Si:P. There are two central results that emerge from the measurements. First, even for $n > n_c$, there are isolated clusters of impurities which contain localized states, some of which are magnetic. This is to be expected on purely statistical grounds, and it is consistent with the notion of coexisting localized and extended states. The separation of the two is quantitatively more difficult in the magnetic case, because of the range of size available to the clusters, which is T dependent because the critical size is defined by the cluster energy-level spacing being of order kT. Since the level spacing decreases like N^{-1} for clusters containing N impurities, and $\langle N \rangle$ increases for n increasing in the neighborhood of n_c, this is a complex question which requires analysis beyond the limits of the usually adopted pair approximation.

There is a second feature of the data which is more important because it transcends the numerical details of curve fitting. This second feature of the data is the absence of critical behavior in the specific heat[22,29] of uncompensated samples for n near n_c. Classical scaling models predict that some kind of analytic singularities in the specific heat and magnetic susceptibility must occur when a transition occurs in the conductivity (see, for example, Ref. 12), but none are observed. Instead only a gradual formation of Schottky anomalies associated with localized states appears as n decreases through n_c.

This second feature is perfectly understandable in the

context of the quantum theory of measurement. The $T=0$ limit of the conductivity measured with an electric field depends on extended states which are separated from the localized states only when such a field is applied. In the absence of such a field, the localized and extended states remains mixed and so no singularity appears in the specific heat or the magnetic susceptibility. This is a characteristically quantum-mechanical effect, and so it is not surprising that classical scaling theories[7,11,12] cannot explain it.

VII. CONCLUSIONS

The present analysis differs from previous work[11,12] primarily in its emphasis on the quantum nature of the TPR transition and its significance in the context of the quantum theory of measurement.[19,20] The present point of view, which is fully consistent with experiment,[22,29] is that the TPR transition is not a classical phase transition, and that it cannot be interpreted consistently by using classical scaling methods.[7]

[1]N. F. Mott, Proc. Phys. Soc. London Sect. A **62**, 416 (1949).

[2]N. F. Mott, Can. J. Phys. **34**, 1356 (1956).

[3]N. F. Mott, Philos. Mag. **6**, 287 (1961); **44B**, 265 (1981).

[4]P. W. Anderson, Phys. Rev. **109**, 1492 (1958).

[5]G. A. Thomas, M. Paalanen, and T. F. Rosenbaum, Phys. Rev. B **27**, 3897 (1983).

[6]M. H. Cohen, J. Non-Cryst. Sol. **4**, 391 (1970); Can. J. Chem. **55**, 1906 (1977).

[7]E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).

[8]W. Götze, J. Phys. C **12**, 1279 (1979); Philos. Mag. B **43**, 219 (1981).

[9]F. J. Wegner, Z. Phys. **25**, 327 (1976).

[10]G. A. Thomas, Y. Ootuka, S. Katsumoto, S. Kobayashi, and W. Sasaki, Phys. Rev. B **25**, 4288 (1982). Larger values of the compensation ratio K in Si(P,B) have been studied by M. J. Hirsch, U. Thomanschefsky, and D. F. Holcomb, Phys. Rev. B **37**, 8257 (1988). They find for $0.2<K<0.5$ that $\alpha=0.9(1)$ with measurements extending down to $T=0.2$ K. This is consistent with the low crossover compensation $K_c\sim 0.05$ assumed here.

[11]C. Castellani, C. Dicastro, G. Kotliar, and P. A. Lee, Phys. Rev. Lett. **56**, 1179 (1986).

[12]C. Castellani, G. Kotliar, and P. A. Lee, Phys. Rev. Lett. **59**, 323 (1987).

[13]J. C. Phillips, Philos. Mag. **47**, 407 (1983); Philos. Mag. B **58**, 361 (1988); Solid State Commun. **47**, 191 (1983).

[14]J. C. Phillips, Europhys. Lett. **14**, 367 (1991).

[15]J. C. Phillips, Phys. Rev. B **43**, 8679 (1991); H. F. Jang, G. Cripps, and T. Timusk, *ibid.* **41**, 5152 (1990).

[16]P. B. Allen and J. L. Feldman, Phys. Rev. Lett. **62**, 645 (1989).

[17]H. B. Shore and J. W. Halley, Phys. Rev. Lett. **66**, 205 (1991).

[18]P. J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966), p. 3.

[19]A. Einstein, B. Podolsky, and N. Rosen, Phys. Rev. **47**, 777 (1935); U. Fano, Rev. Mod. Phys. **55**, 855 (1983).

[20]N. Bohr, Phys. Rev. **48**, 696 (1935).

[21]N. Kobayashi, S. Ikehata, S. Kobayashi, and W. Sasaki, Solid State Commun. **32**, 1147 (1979).

[22]M. Lakner and H. v. Löhneysen, Phys. Rev. Lett. **63**, 648 (1989); H. v. Löhneysen, Adv. Solid State Phys. **30**, 95 (1990).

[23]M. Tiles, *The Philosophy of Set Theory* (Blackwell, Oxford, 1989), pp. 118–134.

[24]S.-Y. T. Lin and Y.-F. Lin, *Set Theory with Applications* (Mariner, Tampa, 1981), pp. 167 and 168.

[25]N. F. Mott, J. Non-Cryst. Solids **1**, 1 (1968); N. F. Mott and E. A. Davis, *Electronic Processes in Non-Crystalline Materials* (Clarendon, Oxford, 1979), p. 34ff.

[26]Y. Imry and S.-K. Ma, Phys. Rev. Lett. **35**, 1399 (1975).

[27]S. Kirkpatrick, Rev. Mod. Phys. **45**, 574 (1973).

[28]L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, London, 1958), pp. 354 and 359.

[29]N. Kobayashi, S. Ikehata, S. Kobayashi, and W. Sasaki, Solid State Commun. **24**, 67 (1977).